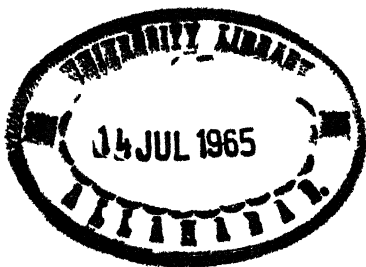
readings in **Experimental Industrial Psychology**

readings in Experimental Industrial Psychology

edited by MILTON L BLUM, Ph D
Associate Professor of Psychology and
Subchairman of the Department of Psychology
School of Civic and Business Administration
of the City College of New York

PRENTICE-HALL, INC, NEW YORK 1952

TO

Donald G Paterson whose high standards
for publication and research have contrib
uted immensely to the development of in
dustrial psychology

Preface

Many instructors prefer to assign readings supplemental to a text Some, not agreeing with the systematic presentations in the various texts, prefer to use assigned readings instead of text assignments Such reading assignments, however, are often difficult because of limited library space and insufficient copies of various periodicals As a result, books on readings in various divisions of psychology fill an obvious need

The major objectives in selecting articles for this volume have been to emphasize the importance of gathering objective data, and to demonstrate that industrial psychology is primarily experimental Appropriate articles have been selected to emphasize that psychologists have devised relevant but varied experimental methods for gathering data on the multitude of problems concerning man and his work

I have chosen the name Experimental Industrial Psychology to represent collectively all the material The experimental method in industrial psychology is only now being recognized While many agree upon the place of the experimental method, there is not as much agreement on what the name should be or whether the field should be separated entirely, or in part, from industrial psychology By way of illustration, applied experimental psychology, biomechanics, human engineering, and applied psychophysics are some of the names now assigned by those who prefer separation Some prefer to limit the area of study to the machine and man Here, it has been thought advisable to broaden the scope and to include man and the various aspects of his work Industrial psychology and the experimental aspect are considered as conjoined

This volume presents five parts of industrial psychology, namely personnel problems, human relations, engineering psychology, consumer and advertising, and newer concepts The first three chapters include material revealing the industrial psychologist as a personnel man The next three chapters concern the industrial psychologist as a promoter of effective human relations and deal with motivation, conflict, and production Part Three is organized into four chapters and is concerned with the relationship of the psychologist to the engineer Part Four is a fertile field for research and shows the contribution of the psychologist to the businessman interested in product distribution The last section describes the psychologist as a researcher always testing and trying new ideas

No attempt has been made to cover all the subjects that could have been included in each of the parts of this book Many topics have been entirely omitted To name just a few one would have to list the interview, job and employee evaluation, accident proneness, supervision, public opinion and market research polling, time and motion study, and work space arrangement

The articles chosen are primarily recent This does not mean that earlier work was not regarded as equally important This merely means that this book of

readings refers to contemporary contributions rather than the classic antecedents because of space limitations

In the section on engineering psychology, especially, reference is made to contributions stemming from work in the armed forces Such material has been included since this research is not always clearly distinguishable from industrial research Consulting organizations and universities are likely to have contracts for research in both fields and, indeed, the former definitely have industrial implications and applications

To highlight the major problems of research in each topic, each section and chapter begins with brief introductory remarks No attempt has been made to evaluate critically the articles included Systematization of the research according to method, subjects employed, criteria, and implications of the conclusions may or may not be classroom topics, depending upon the judgment of the potential user of the book

The volume is intended for various uses It can suggest experimental design and procedures in considering research projects It can serve as assigned readings for such courses as industrial psychology, applied psychology, or experimental psychology It is hoped that it will encourage offering courses in experimental industrial psychology

The volume owes its value to the work of the original writers whose material has been included I wish to take this opportunity to thank them collectively for the generosity and cooperation offered The original publishers were very kind in granting reprint permissions and their lack of selfishness is deeply appreciated Specific acknowledgments to authors and sources of publication are given with the respective selections

I am indebted to many persons for their helpful suggestions and their reading of parts or the entire manuscript I wish to specifically thank Dr Jesse Orlansky, Dr Paul M Fitts, Dr James Jenkins, Dr Wallace Russell, Dr K V Smith, Dr Benjamin Balinsky, Dr Max Smith, Lt Colonel Walter F Grether, and Professor Donald G Paterson Mrs Irma Schneider and Mrs Winifred Bullock have been tremendously helpful through their intelligent secretarial assistance The kind encouragement from my wife and children has been of vital importance to me in this undertaking

MILTON L BLUM

Contents

PART ONE PERSONNEL PROBLEMS

PART TWO HUMAN RELATIONS

PART FIVE NEWER CONCEPTS

PART ONE

Personnel Problems

Selecting and training employees are ordinarily considered the major role of personnel departments In industry, psychologists usually but not always devote their attention to such tasks Depending upon company policy, however, any of many differently trained persons may hold such positions as director of training or personnel Many firms have a policy of promotion from within and in such instances some executives may have the benefit of experience but not of any formal training for the position

Three chapters have been included in this section to illustrate the kind of re search that psychologists are likely to conduct Psychological Tests, The Application Blank, and Training have been selected from among many chapters that might have been included Each chapter is intended to point to specific techniques that are used in solving some of the problems related to the area

Chapter I

PSYCHOLOGICAL TESTS AND EMPLOYEE SELECTION

Psychological testing in industry is rather widespread and is the area that affords the psychologist his most likely entree to businessmen As the references in other chapters will establish, however, the psychologist is capable of performing many other tasks

Psychological testing has now become so acceptable that many who lack qualifications test indiscriminately without understanding the difference between correct and incorrect application Some test distributors recognize this and have established rules for restricting the sales of tests They are to be commended insofar as they obviously are more concerned with correct usage than with sales

A major difficulty connected with psychological testing in industry is that testing cannot correctly be installed without experimentation A psychological test is not merely a series of questions The test must have certain characteristics, the most important of which is validity There must be a relationship between test results and successful job performance In order to establish this relationship, one needs much more than a test and a measure of job performance One must really conduct an experiment Different groups of subjects with known characteristics must be measured to insure that the results obtained will not be spurious Controls must be introduced and exacting care must be exercised in procedure as well

as in data analysis Ultimate accuracy depends upon cross-validation of the system of checking results with a new group, in addition to the experimental and control groups used to obtain the results

Unfortunately, as tests and testing methods gain wider acceptance, services are sometimes sold that cannot possibly have value The businessman who buys such services is "taken" and when he discovers this he blames tests in general rather than his poor judgment in particular The easiest way to avoid such an error is to consult a competent, professional psychologist A diplomate in industrial psychology (American Board of Examiners in Professional Psychology) is likely to be such a person Fellows or Associates in the Business and Industrial Division of the American Psychological Association also have adequate professional qualifications Such individuals are likely to have had their enthusiasm tempered by experience Accordingly, for problems of testing, they are likely to have more mature judgment and be more responsible than others without such qualifications They are less likely to make unsubstantiated claims and will strive to be more scientifically accurate in reporting results

Much prior knowledge is necessary in selecting tests for industrial usage and in conducting a thoroughly accurate study Different tests are required, for example, depending upon whether one is hiring experienced or inexperienced applicants Also, subjects are rarely so naive that they cannot modify their answers according to the way they think their answers "should be" rather than "are" Introducing certain controls can minimize this tendency The manner of administering a test requires more than merely reading from a manual of directions, and interpreting test results depends to a larger extent upon the relative performance of the group being studied rather than upon an absolute but arbitrary performance imagined to be desirable

Selecting a few studies to illustrate the complete range of problems investigated in relation to selecting employees with the aid of psychological tests is practically impossible The range of psychological tests includes such diverse areas as intelligence, aptitude, achievement, interest, and personality Both pencil and paper as well as apparatus tests have been designed and, in addition, the manner in which the test is administered also varies from individual testing to group testing

Psychological tests have been used as aids in selecting employees for a wide range of jobs Factory workers of assorted types, clerical workers, and executive personnel are merely three broad categories serving as illustrations

Although the readings in this book have been selected primarily to emphasize and illustrate experimentation, the study by Margaret Jones should be included This report is a result of a critical review of over 2,000 references on employee selection It posits high and thoroughly basic requirements for such studies It also refers to eight studies which have met the criteria of adequacy in both experimental design and report Jones briefly lists the five requirements of a "good report" which, needless to say, should form an outline for all future studies in this area

The Blum and Candee study attempted to select packers with the aid of performance tests It is cited since Jones included it in her list of eight studies selected as meeting all criteria of adequacy

The Hay study is concerned with a clerical job and the tests used are primarily of the pencil and paper, and group administered variety The study not only reports initial success but also indicates the results of a ten year follow-up

The Sartain article is included for two reasons First, it is one of the few studies meeting Jones' requirements, and second, it illustrates that even a professionally competent psychologist using his best judgment in selecting tests may find that such tests have little or no predictive value as aids in selecting employees for a specific job The test battery included a wide range of types of tests and the selection problem was on the supervisory job level

This study illustrates the principle that negative results may be reported and have value In fact, it is more desirable to report negative results than to confuse and bewilder by exaggeration or misinterpretation

The Stromberg study is included because it breeds controversy It takes a stand in favor of "any battery of tests" The editors' footnote in the article is most likely an elaboration or clarification of an otherwise 'out-on a-limb" viewpoint The Stromberg report finds that a testing program draws better applicants The battery included tests of intelligence, personality adjustment, and aptitude and the job was the production type of factory work

Taft reports a more recently developed and complex type of testing, namely the group situation examination This technique is an outgrowth of experiments conducted during World War II and shows promise as an aid in selecting professional, managerial, and other higher level employees Included in the Taft study was the group Rorschach, a projective-type test used more widely in clinical psychology than in industrial psychology

The Adequacy of Employee Selection Reports *

MARGARET HUBBARD JONES

An examination of well over 2,100 references from the obtainable world literature on employee selection has permitted an analysis of practices in both experimental design and report the results of which are perhaps surprising These references cover the period 1906 to 1948, within the ability of the author to locate the references and within the ability of reference librarians to locate copies of unusual and foreign periodicals and monographs There are further limitations to the data to be analyzed here which were imposed by the primary purpose of the literature search That aim was the compilation of abstracts of employee selection reports which should contain the actual data presented, together with sufficient information to enable the reader to evaluate the study without referring to the original report The work was prompted by the difficulty experienced in this particular field in locating the widely scattered references (we had reference to more than 300 separately titled periodicals—and many volumes of each one—as well as books and monographs) and by the fact that most industrial psychologists do not have the time or facilities necessarv to review this literature These abstracts appear elsewhere (2)

Due to the large volume of material in this field and to the fact that many articles which seem important contain virtually no information, it did not seem economic to abstract all possible references, and the survey is thus limited to those studies which can be evaluated those in which relatively complete validation data are presented, together with specific tests used, N

* Reprinted from *Journal of Applied Psychology* Vol 34, No 4, August 1950

and job studied Further, since we were interested in selection of employees for industrial concerns we have also excluded the special fields of selection for the armed forces and pilot training as posing special problems It has by now become abundantly obvious that seemingly slight changes in working conditions, incentives and parent population, to mention the more obvious factors, may result in the failure of even a carefully executed selection program In view of this, it seemed wiser to exclude those studies in which the criterion was school grades or teachers' ratings, and those carried out on military personnel even where the jobs are similar to civilian jobs, because of the differences in motivation and working conditions After we have eliminated the reports which are, by our definition, special problems and those which are so inadequately presented or executed that they cannot even be evaluated by the reader (a large proportion of the total) there remain 427 reports, or 20 per cent of the total number of references In this analysis we shall be concerned entirely with these 427 reports which can be evaluated Since many articles report results on diverse jobs very often with different tests and different statistical procedures we have at times referred to the number of separately treated groups rather than to the number of titles and have endeavored to make the distinction clear wherever the former occurs

These 427 references are largely American—about 80 per cent—both because of the greater availability of privately published and unpublished material of American origin and because of the larger total volume of articles It does not appear that the percentage of acceptable articles is much larger for American work than for that of any other country

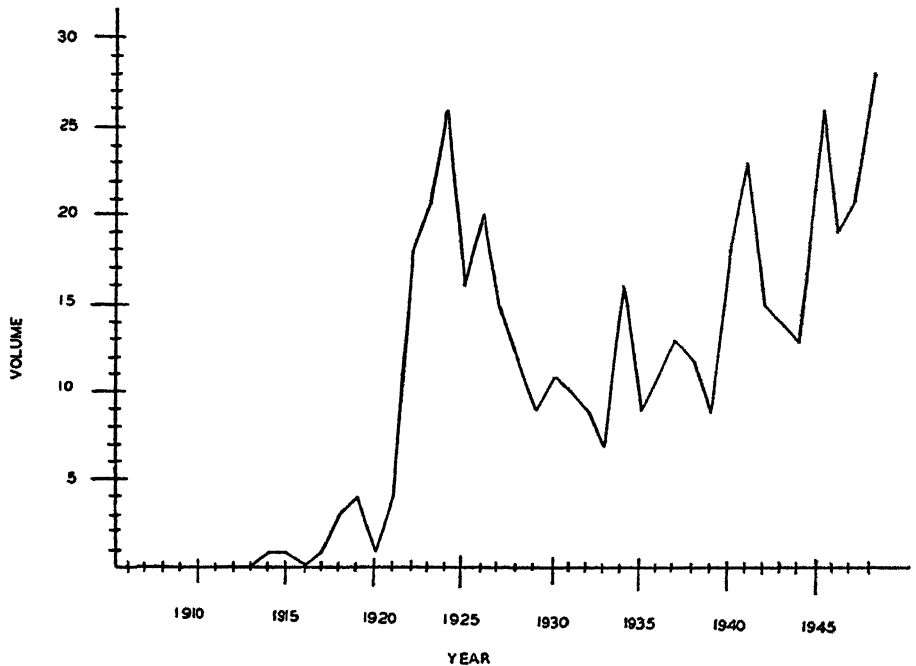



FIGURE 1 1 *Number of employee selection reports by year*

The volume of acceptable articles is shown by year in Figure 1 1 The slow rise to a peak after World War I, followed by a decline through the depression era is not unexpected Whether this can be entirely attributed to overselling of testing, as is usually done, or whether it does not also reflect to some extent the general decline in business activity is a debatable point The annual volume of articles reached a

high again in 1941, as business was recov ering, fell off, understandably, during the war years, and has now reached an all time high Whether it will remain high probably depends partly on the quality of current work and partly on the general level of industrial activity

The ten jobs which have been most frequently studied are as follows Sales men, 75, Clerical Workers, 60, Teachers, 49, Assemblers, 23, Executives, 23, In spectors, 23, Supervisors, 21, Typists, 17, Stenographers, 14, and Machinists, 9 This order does not necessarily reflect either the importance of the job or the difficulty of selecting good workers Among the jobs which appear to have been acceptably re ported but once are brick layers, grocers, scientists and deans of women As can be seen, salesmen lead the list This, of course, includes salesmen of all sorts and many of the jobs are quite different The same comment applies to the other categories There is usually no way of determining from the published reports whether the jobs in two studies are comparable As Ghiselli has shown, there is an astounding range of reported validity coefficients for the same general type of test in any broad occupational classification (3) For cler- ical occupations he found the range to be over 90 There are many reasons for this state of affairs but one which has perhaps not been sufficiently emphasized is the lack of adequate job description in published reports It is now fairly generally recog nized that a selection program is very much situation bound but the corollary, that this requires precise job description, is not currently practiced

Let us now examine in more detail the 427 reports which represent the cream of the crop The number of subjects used in the investigation is an important factor in determining the predictive value of the results Except in the majority of cases using less than 20 subjects, the N in and of itself does not tell the whole story Much depends upon how the data are treated and whether or not the total pop ulation of employees on a particular job, or a representative sample thereof, was used Nevertheless, it is instructive to an alyze the trend in this respect The results of the analysis by number of subjects is as follows Less than 10, 17, 10–19, 97, 20–29, 93, 30–49, 129, 50–99, 188, and 100 and above, 257 Even more unexpected than the number of groups with small N is the number with 50 or more and the 257 groups containing 100 or more subjects The latter are by and large the more recent studies and the trend is encouraging

Statistical Techniques

An analysis of the statistical techniques used for presentation of the results of val idation procedures is interesting, but again the particular statistic used does not guar antee adequacy of treatment because the assumptions governing its use may not have been met and the statistic best suited to a given problem may not have been chosen Table 1 1 shows the frequency with which various measures are used Correlational techniques are the most popular, account ing for 285 out of 525 separately treated groups Of these only 172 give measures of significance and although they can be calculated from the data provided, it is safer, considering the heterogeneous nature of the audience in this field, to present the standard errors along with the coefficients of correlation Furthermore, the author has the real responsibility for complete pres entation of all the statistics necessary to an interpretation of his investigation Oc casionally one even finds an author con cluding that the correlations reported have clearly shown a relationship between test scores and criterion whereas actual calcu

TABLE 1 1

Statistical Measures Used for Validation

Measure of Correlation	*Number of Groups So Treated*
r	136
Rho	94
R	35
r_{bi}	7
tetrachoric	8
other	5
Total	285
Group Comparison	185
Inadequate Treatment	55

lation shows the correlations to be not significantly different from zero This practice is general and no single individual should shoulder the blame for it

Group comparisons of various sorts account for 185 cases, but of these only 28 include measures of the significance of group differences (although sometimes such measures could be calculated by the reader) Group comparisons may take such forms as differences in mean test scores between the upper and lower 50 per cent of employees as judged by the criterion, or average test scores for groups judged best, average and poorest by their supervisors (many times without N or sigma for each group being indicated), or per cent of those scoring within certain limits who were judged good as against the per cent who were judged poor, etc Only occasionally are critical ratios or t ratios or similar measures included The importance of testing results for significance cannot be overemphasized In the case of group comparisons it is more serious than where correlation is used because in most cases there are not sufficient data to enable the reader to perform the proper calculations for himself In one particular case, where the author concluded that his tests were efficacious for selection but neglected to supply any measure of the significance of the differences found between groups, calculation of the significance of percentage differences (the only data available) showed the differences to be exceedingly insignificant

In 55 cases there is incomplete statistical analysis In a few cases the raw data are presented with no summary statistics In many instances we find the results expressed only as per cent agreement" between test scores and criterion scores, or a brief statement that a critical score of a certain magnitude would have eliminated a given per cent of the poor group and ordinarily a smaller per cent—although we do not know that it is a reliably smaller per cent—of the good group In 5 cases the authors are content to present graphs alone, sometimes with the differences very much exaggerated by the scale and baseline chosen

One gains the impression that many times, even where adequate statistics are used, the basic requirements for their use have not been met[1] One should be able to assume, for example, that when an r is reported the conditions for its proper use have been met, but in view of the inadequacy of many of the statistical treatments one cannot always so assume A more obvious criticism of many studies is the manner in which subjects are selected In spite of the fact that the assumptions underlying many of the statistics used require a reasonably random sample, biased rather than random sampling seems to be the rule It is a common procedure to select certain employees to serve as subjects but rarely are we given any information which indicates that the sample was a selected one or how it was selected A frequent practice is the artificial creation of a heterogeneous experimental group by the use of only extreme employee groups (the upper and lower 25 per cent, for example) a practice which may spuriously raise the validity coefficients

A point too often overlooked is that a selection program is intended to select among *applicants,* not among employees, and the two groups are not identical (cf 5, 7, 10) "Natural selection" on the job—the survival of the fittest—operates to make the employee group more homogeneous than the applicant group This may spuriously lower validity coefficients and change critical scores Further, the employee group will often not show a normal distribution in a trait which is highly correlated with ability to produce on the job and since in most industrial situations it will be impossible to correct for this error, the usefulness of the employee group as a basis for a selection program is further limited The best

[1] The criticism of the use of statistical methods in research on the Rorschach Test by Cronbach (1) may be applied in part to employee selection research even where other tests are used Especially to be noted are his criticisms of the selection for emphasis of a few significant" differences from among many insignificant ones, whether the comparisons are explicitly made or merely implied, and his insistence on the use of a second independent sample so that chance variations in test scores will not be given undue weight

practical solution to both problems—that of bias in sampling and that of restriction of range in employee groups—seems to be the use of two groups first, a randomly selected employee group as a trial group, for reasons of economy, and second, an unselected applicant group as a follow up group, to discover whether or not the selection program will select among applicants as well as among employees This solution, the use of separate groups, has the further advantage of permitting a pragmatic estimate of the shrinkage in multiple correlation This is a real advantage An example is Selover s two samples of clerical workers (N = 193 and 85, respectively) which yielded multiple correlations for 4 tests and criterion of 41 and 33, respectively (9) An instructive example of the danger involved in putting one s faith in a single small sample, particularly when that sample has been used to develop a scoring procedure, is given by Kurtz (4) Here a scoring technique for the Rorschach Test was developed which classified correctly 79 out of 80 sales managers This was so impressive to many people concerned that they were prepared to start using it as a selection device immediately A follow up on a second sample yielded a validity coefficient of 02[1]

The criterion is, of course, a question of utmost importance but we cannot discuss all the ramifications of the problem here Entirely aside from the question of the applicability of the criterion as a real measure of job success—the validity of the criterion—we find a problem in the reliability of the criterion Only 95 reports, or 22 per cent of the 427 acceptable reports, make some attempt to include measures of the reliability of the criterion, and yet it has a profound influence upon the results of the validation procedures Of course, low reliability will not give spuriously high validities—rather the opposite—but many studies appear to lead to the conclusion that certain tests are worthless in a given situation, whereas low criterion reliability may actually be masking a significant relationship Surely, if a study is worth doing at all, the reliability of the criterion should be ascertained

Another difficulty in connection with the criterion is the operation of external influences such as age, experience and length of time on the job[2] Unless cognizance is taken of these variables the results are difficult to interpret, to say the least, and few studies control any of these factors For example, it is easy to see how age may be predictive of job success if age is influencing the criterion either in its own right or operating through length of service, and yet most studies lump together not only all age groups but employees with widely differing lengths of service On the other hand, if age or length of service is influencing the criterion a significant relationship with test scores may be masked One often suspects a further contaminating factor when ratings are used if test scores are not kept strictly confidential until after ratings are made The facts seem to indicate that more attention must be paid to proper experimental design if the results of selection studies are to be useful

How many reports, then, meet all criteria of adequacy in both experimental design and report? We find that 46 out of the 427 originally selected contain no second or follow up group but are acceptable in all other respects, such as sufficiently large N, adequate and complete statistical presentation throughout, etc We further find that 17 are adequate in all respects except that no measure of the reliability of the criterion is presented Finally, if we count the total number of reports which are satisfactory in all respects we discover only eight or 4 per cent of the 2100 references with which we started These eight studies are as follows

1 Bellows R M, "Studies of Clerical Workers, Chap VIII in Stead, W H, Shartle, C L, et al, *Occupational Counseling Techniques* New York American Book Co 1940, ix + 273, pp 144–146 (Study of coding clerks)
2 Blum, M and Candee, B, The Selection of Department Store Packers and Wrappers With the Aid of Certain

[2] An attempt to compensate for bias in the direction of longer service may be found in a recent study by Rundquist and Bittner (8), and McMurry and Johnson attempted to secure ratings on their subjects after approximately equal time on the job (6)

Psychological Tests,' *Journal of Applied Psychology* 1941, Vol 25, 76–85

3 Guilford, J P and Comrey, A L, Prediction of Proficiency of Administrative Personnel from Personal History Data, *Educational Psychological Measurement* 1948, Vol 8, 281–296

4 Holliday, F, The Relation Between Psychological Test Scores and Subsequent Proficiency of Apprentices in the Engineering Industry, *Occupational Psychology* London, 1943, Vol 17, 168–185

5 Otis J L, Endler, O L, and Kolbe L E, Data Analysis Methods ' Chap VII in Stead, W H, Shartle, C L, et al, *Occupational Counseling Techniques* New York American Book Co 1940 ix + 273, pp 113–136 (Study of department store salespersons)

6 Rundquist, E A and Bittner, R H, 'Using Ratings to Validate Personnel Instruments a Study in Method,' *Personnel Psychology* 1948, Vol 1, 163–183

7 Sartain, A Q, Relation Between Scores on Certain Standard Tests and Supervisory Success in an Aircraft Factory, *Journal of Applied Psychology* 1946 Vol 30, 328–332

8 Selover, R B, 'The Development and Validation of a Battery of Tests for the Selection of Clerical Workers, *American Psychologist* 1948, Vol 3, 291–292 (abstract), and personal communication

It is not intended to imply that these studies found highly predictive test batteries, but merely that the technique was adequate Conclusive negative findings are important and are too frequently ignored or even suppressed

In conclusion, let me emphasize two points First, the actual work done by industrial psychologists is not as bad as would appear from this analysis, and the trend is definitely toward more complete and careful design and execution In many cases our criticisms apply to the reports, not necessarily to the studies themselves More care should be taken in the preparation of reports so that all relevant information is available to the reader

Requirements of a 'Good Report"

Perhaps a summary of the items one weary abstractor would like to see made explicit would be in order

1 Detailed job description, with each group treated separately

2 Complete description of the sample N (sufficiently large), what proportion of the total population this represents and how selected, factors involved in hiring, age, length of time on the job (preferably with widely differing employees treated as separate groups), and total experience in jobs of similar nature, use of two samples, one an applicant group

3 Exact test titles, when in the employment experience the tests were administered, whether the tests were a factor in hiring, where the tests were given, under what conditions and incentives the tests were given, reliabilities of tests with comparable groups

4 Detailed description of the criterion, length of time on the job when the criterion measure was applied (with widely differing employees treated as separate groups), reliability of the criterion, some discussion of the validity of the criterion selected, if ratings are used, some estimate of the amount of contact the rater has with the employee, if production records are used, the duration of the period and whether there were any unusual factors operating at that time

5 Adequate statistical treatment, with assurance that the assumptions governing the use of the given measures have been met, and actual report of the numerical results, together with an appropriate measure of significance

This may seem like a large order, but many adequately executed studies already reported could have included most of the items since it is obvious from certain remarks that the author must have taken them into consideration In view of the untrustworthiness of many reports these items should be made explicit

A final point concerns those studies done by inadequately trained personnel There are many of these and they are quite useless They point to the ultimate desirability of some method of identification of properly qualified personnel for employee selection programs

SUMMARY

A survey of more than 2,100 references on employee selection in industry has re vealed that only 427 contain sufficient in formation to permit evaluation of the study These 427 reports are analyzed in terms of annual volume, jobs most frequently investigated, statistics used in pres entation of validity, number of subjects and general adequacy of design This analysis reveals that many of these studies are inadequate to permit drawing conclusions as to the efficacy of the selection procedures employed Factors which influence results but are difficult to evaluate from reports as they are usually published are discussed Some recommendations for items to be included in reports of employee selection programs are presented

REFERENCES

1 Cronbach, L J Statistical Methods Applied to Rorschach Scores, *Psycho logical Bulletin,* 1949, Vol 46, 393–429

2 Dorcus, R M and Jones, M H, *Hand book of Employee Selection* New York McGraw Hill Book Company, Inc 1950

3 Ghiselli, E E, The Validity of Com monly Employed Occupational Tests' *University of California Publications in Psychology* 1949, Vol 5 (9), 253–288

4 Kurtz, A K, A Research Test of the Rorschach Test,' *Personnel Psychol ogy* 1948 Vol 1, 41–51

5 MacMillan, M H and Rothe, H F, Additional Distributions of Test Scores of Industrial Employees and Appli cants, *Journal of Applied Psychology* 1948, Vol 32, 270–274

6 McMurry, R N and Johnson D L Development of Instruments for Se lecting and Placing Factory Employees, *Advanced Management* 1945, Vol 10, 113–120

7 Rothe H F, 'Distribution of Test Scores of Industrial Employees and Ap plicants *Journal of Applied Psychol ogy* 1947, Vol 31, 480–483

8 Rundquist E A and Bittner, R H, Using Ratings to Validate Personnel Instruments a Study in Method, *Per sonnel Psychology* 1948 Vol 1, 163–183

9 Selover, R B, The Development and Validation of a Battery of Tests for the Selection of Clerical Workers, *Amer ican Psychologist* 1948, Vol 3, 291–292, and personal communication

10 Stromberg E L, "Testing Programs Draw Better Applicants,' *Personnel Psychology* 1948, Vol 1, 21–29

The Selection of Department Store Packers and Wrappers with the Aid of Certain Psychological Tests *

MILTON L BLUM

and

BEATRICE CANDEE

A large department store hiring many packers and wrappers for the Christmas season asked the New York State Employ ment Service to investigate the use of tests for this purpose They agreed to have the employees who had already been engaged on such jobs tested so that the validity of the test results could be checked with this group as well as with inexperienced workers

Fifty three permanent employees were given the O Connor Finger Dexterity Test the Minnesota Placing Test and the Min nesota Turning Test All tests were ad

* Reprinted from *Journal of Applied Psychology,* Vol 25, No 1, February 1941

ministered by an experienced examiner The employees were tested in a specially designated room at the store They were informed by both management and the psychologist that the test results were in no way to be considered as a check on their ability, but were being given to see if such tests could select future employees Apparently there were no unfavorable emo tional attitudes on the part of the subjects

The second group of subjects consisted of one hundred thirty people who were given the three tests at the New York State Employment Service Of this number, thir-teen were eliminated on the basis of ex-ceptionally poor test results since the store felt that this should be a service as well as a research project The employment service ultimately made ninety two referrals to the store All but three, who for some reason did not meet the requirements of the store, were hired This group will hereafter be referred to as the seasonal employment group

The permanent employees tested worked as packers and wrappers, relief cashiers, cashiers, and assorters Production records were available for the thirty six employees engaged on the first two mentioned jobs only Supervisor s ratings were available for thirty eight of these employees The average production for a six months period for the relief cashiers was 12,691 units The packers and wrappers achieved an average production record of 25,759 units This meant that the different jobs offered different opportunities Of the eighty nine seasonal employees hired, only 52 subjects were employed on jobs which could be measured on a production basis Super visor's ratings were available for forty four of these subjects

Table 2 1 presents the average test score for each group on the various tests The test results for the various employee groups show that the new workers when selected had slightly inferior test scores than the permanent employees Only on the Placing Test, however, is the difference between the averages statistically significant In other words, with the elimination of only one tenth of their referrals on the basis of tests, the interviewers of the employment service had selected a group only slightly inferior in test performance to the perma nent employees on these jobs This in itself is interesting Four years before this when using the Finger Dexterity to select similar workers in another store we found that in five weeks time, interviewers developed a 97 per cent accuracy in selecting applicants to pass a critical score on the test

The intercorrelations of the tests based upon a sampling of one hundred thirty fe males seeking the positions open were as follows

r between Finger Dexterity and Placing Test was + 416 ± 07

r between Finger Dexterity and Turning Test was + 335 ± 07

r between Placing and Turning Test was + 551 ± 06

The attaining of an acceptable criterion of success was as usual, a most difficult task The most tenable criterion was the actual production record of each employee During the month of December all em ployees are probably working closer to maximum ability than at any other time of the year It was, therefore, decided to consider the average daily number of pack ages wrapped during this month, as the criterion

TABLE 2 1

Average Score on Tests for Seasonal and Permanent Employee Groups

	No	*F D aver*	*F D σ*	*Plac aver*	*Plac σ*	*Turn aver*	*Turn σ*
Seasonals hired as wrappers	27	7′ 48	46	220′	16″	175″	19′
Seasonals hired as packers	10	7 41	43″	225′	11″	174″	13″
Seasonals hired as wrap cashiers	15	7 56	57″	225″	9″	178″	13
Permanently employed relief cashiers	11	7 24	52″–	205″	15″	166″	17″
Permanently employed wrap cashiers	25	7′ 20	47″	201″	12″	167″	23″

This criterion is acceptably reliable The reliability coefficient using the split half technique, when average production for the first half of the month of December was compared with the average production of the second half, was + 88 Even when different seasons of the year are compared, a correlation (rank difference method) of + 70 was obtained between the average daily production record for the group of thirty six permanent employees over a six month period and the average daily production record for those employees during the month of December

The average daily production in December for each of the groups investigated is presented in Table 2 2

While the 10 per cent of the distribution which was eliminated would probably reduce the correlation coefficient, it is not likely that the relations would have become meaningful if it were included in the case of either the Turning or the Finger Dexterity

It is interesting that both the Placing and Turning give higher correlations with the new workers than with the permanent group, which apparently indicates that even the slight predictive value possessed by the tests is reduced by experience on the job

The relation existing between the Finger Dexterity Test and production record for both groups is zero This result is rather surprising in view of the fact that the Finger Dexterity Test has had much prestige among department stores in New York City in the selection of packers

TABLE 2 2

Average Daily Production During December

	Average daily production	σ
Seasonals employed as wrappers	102	27
Seasonals employed as packers	111	20
Seasonals employed as wrap cashiers	79	17
Permanently employed wrap cashiers	172	76
Permanently employed relief cashiers	91	43

The permanent employees seem to be clearly superior on the basis of production records On the jobs which are directly comparable, those of wrapper cashier, their performance is more than twice as good This difference between the groups is much greater than shows on the test scores and might indicate either that experience largely determines the better performance of the permanent workers or that these particular tests do not get at the abilities involved An additional factor probably is that the permanent workers are given preference for the more productive jobs

Table 2 3 presents correlations existing between the criterion, production records, and the various test results

The Placing Test yields better results However, although the correlations with the Placing Test are as high as those frequently reported, they would in no way justify the selection of workers on the basis of this test alone

The multiple correlation between production records of the seasonal employees and all three test scores was + 38 and + 24 between production records of the permanent employees and test results This indicates that the 3 tests together give only slightly better results than the Placing

TABLE 2 3

Correlation between Production Record and Test Results

	Plac	*Turn*	*F D*
Seasonal employees production record	35	27	02
Permanent employees production record	21	06	08

alone A relationship is present for the seasonal workers, which is reduced by experience as in the case of the single tests However, even with the tests combined the correlations between production records and test results were too low to be valuable for prediction in individual cases It was decided to compare the various groups when divided according to high production and low production records (High defined as the highest third, low defined as the lowest third) The average score for each of the groups mentioned in Table 2 2 was then computed Naturally, when subdivided to this extent, the numbers of cases in each comparison is small and does not warrant the actual presentation of the average time to complete the test and the standard deviation However, the comparisons were in the expected direction (average time of highest third in production record faster than average time of lowest third in production record) in 11 of the 15 instances Like the correlations, this method would indicate that the tests are not valueless but that the test results are not good enough when individual selection is the problem to be solved However, when mass hiring for seasonal work is necessary, such testing may be of some aid to the personnel department provided it is worth the cost of administering the tests, and providing a serviceable critical score can be found for the specific store

One additional criterion, supervisor ratings, was available The ratings obtained in this study were made by supervisors as a routine procedure They rate all employees with the following consideration in mind

Rate A —all who are to be continued in employment

Rate B —all who are good and should be re hired at the first opportunity available

Rate C —all who are to be re hired for seasonal work

Rate D —all who are inefficient

Here the permanent employees are apparently clearly superior However, these ratings require further explanation because of the policy of the Personnel Department of the store Since most permanent employees are to be continued in employment, and since with very rare exceptions all seasonal workers are laid off after the holidays and rehired later as the opportunity occurs, a ' B" rating for a seasonal worker may easily be equivalent to an ' A' for a permanent employee However, the comparison cannot even be made simply on this basis because the practice of defining an "A' as designating a person to be continued in employment undoubtedly eliminates practically all differences among the permanent employees except the distinction between those who are actually unsatisfactory and those who are acceptable

A comparison of the test scores of the various groups in Table 2 5 shows that the permanent employees with ' A" ratings obtain average scores on all three tests that are superior to the average test scores of all seasonal employees, both those rated C' and those rated "above C The D/σ_{diff} in all cases is greater than three with one exception (The D/σ_{diff} between average F D score for rating of "A' group and Rating of "C" group = 2 7) The D/σ_{diff} between the permanent employed group obtaining an "A" rating and both seasonal groups on the placing test is greater than 7

TABLE 2 4

Distribution of Supervisor s Ratings

	N	*A*	*B*	*C*	*D*
Seasonal employees	44	2%	52%	46%	0%
Permanent employees	38	82%	13%	5%	0%

TABLE 2 5

Comparison of Test Results for Various Groups

Finger dexterity	*N*	*Average*	*Range*	σ	σ
Dept store perm employees rating A (Gr 1)	31	436	366–565	46 8	8 32
Permanent employees below 'A (Gr 2)	7	457	375–584	57 1	21 92
Seasonal employees Above C (Gr 3)	24	475	407–553	41 5	8 47
Seasonal employees Rated C (Gr 4)	20	476	362–624	54 7	12 23
Queens N Y S E S group—(Gr 5)	420	463	315–735	69	3 38
Placing					
Gr 1		201	174–230	13 2	2 40
Gr 2		209	189–231	13 2	5 07
Gr 3		232	192–260	15 5	3 16
Gr 4		226	198–247	10 8	2 44
Minn sample*		228	140–340		
Turning					
Gr 1		167	141–269	23 3	4 23
Gr 2		163	138–189	16 7	6 29
Gr 3		183	135–221	16 1	3 29
Gr 4		185	154–205	15 4	3 44

* B J Dvorak Differential Occupational Ability Patterns, *University of Minnesota Research Institute* Vol 38, 1935

When the permanent employees rated below A are compared with the seasonal group rated above 'C' the D/σ_{diff} is above 3 on the Placing, but 2 8 on the Turning and 7 on the Finger Dexterity These results indicate that the less satis factory permanent employees are better than the "above average' temporary em ployees on the Placing test They tend to be better on the Turning test and possibly also on the Finger Dexterity test This is still further indication that the store ex perience seems to raise the test score to some extent, even though the tests do not predict better workers within either the more or less experienced groups In all comparisons between the two seasonal groups and between the two permanent groups, no statistically reliable differences are found on any tests

When the employees of the store are compared with available control groups, the permanent employees with "A' ratings clearly excel on the Finger Dexterity test a group of 420 women applying for dex terity jobs at the State Emplovment Serv ice, both in average score ($D/\sigma_{diff} = 3$) and in the elimination of the lower range The good seasonal group, however, is no better in average score on either the Plac ing or the Finger Dexterity than the con trols, although about 8 per cent in the lowest range are eliminated There is no way of knowing whether these people in the lowest range of the test scores would also have been satisfactory workers or whether among them would have been found employees so unsatisfactory that they could not be retained through the season Another department store in New York in a study some years ago on the Finger Dex terity found that by using a much higher critical score than this it could have elim inated 10 out of 14 or 71 per cent of the new workers discharged before the season was over In so doing it would also have eliminated 29 out of 47 or 61 per cent of its acceptable workers and one out of three or 33 per cent of its best seasonal em ployees It seems most improbable that we did eliminate all of the poor workers in the lowest 8 per cent since the other study, which had a better test prediction in all ranges, still would have eliminated a very small proportion of its bad workers on the basis which we used It therefore appears that a genuine difference exists in the re sults of the two studies, due to a difference

in the criteria used, in the handling of the workers, in the store policy, or some other factors in the general situation It must also be noted that even the more favorable results of the other study show so high an elimination of acceptable and even excellent workers along with the less desirable ones that the use of such a method of selection by a public employment service would be questionable even though an individual store may adopt a method of this kind if it has a very large labor supply

In our study the complete absence of any workers rated as inefficient even in the seasonal group is to be noted The store had considered the selection of workers so good that it cut a full day from the training period If the selection actually was much better than previous seasons, it must be attributed to other factors than the tests It is recognized that the interest and attitudes fostered by such a study may in themselves bring about more favorable production records [1]

To summarize In this study the satisfactory permanent employees excel less experienced workers in production, in supervisor s ratings and in test scores, the indications being that this is due primarily to experience on the job That experience rather than selection is the chief factor is indicated by the fact that there are no significant differences on the tests between the satisfactory and less satisfactory permanently employed groups and between the two seasonal employed groups but only between the more and less experienced workers, regardless of the supervisor s ratings Also the correlations between the production records and the scores on the tests are noticeably lower in the experienced group Thirdly, while the satisfactory experienced workers excel the general population on the Finger Dexterity and the Placing, the seasonal workers do not excel in either test although the store considered them an exceptionally good group of new employees

On the basis of these results the only justification for the use of these three tests in the selection of department store wrappers and packers, is if the hiring of seasonal workers is viewed as a mass selection with little attention paid to individuals or to other types of information about the applicant If much consideration is to be given to individuals as prospective permanent employees, then individual test scores for any one person should not be considered too seriously The job of packing and wrapping is a relatively simple one and can be learned within a short period of time as is indicated by the fact that the correlation existing between production of inexperienced people and test scores is reduced with experience

This study seems to indicate, as do most others on dexterity tests, that the specific factors in dexterity functions considerably outweigh any more general ones Apparently experience in wrapping does have a slight effect in raising test scores on three different tests and in reducing initial differences among the workers on the tests However, on the job, specific factors seem to outweigh any which exist in common with either of the three tests since none of them will select good workers from the population as a whole, unless they are already experienced Another study now in progress is being conducted in a different department store to check these results

[1] This was clearly shown in the Hawthorne Study, Western Electric Company Elton May, *Human Problems of an Industrial Civilization* New York The Macmillan Company 1933, pp 194

Predicting Success in Machine Bookkeeping *

EDWARD N HAY

Acknowledgments are extended to the following for critical comments on the design of this experiment and in the preparation of the manuscript George K Bennett, Marion A Bills T J Gorham, Roy B Hackman, A K Kurtz and E F Wonderlic The experiment itself was carried out with the assistance of the following, each of whom contributed substantially to the final successful result Arline Mance Blakemore, Paul K Fryer, Wm D Turner and Mary Elizabeth Hemsath Van Newkirk

The problem is to select from among experienced and inexperienced clerical applicants those girls who, after about a year s experience prove to be rapid and accurate bank machine bookkeepers In the past there has been great variation in the amount of satisfactory work produced by different girls in an hour's time For example, the extreme range in the last five years is from the slowest operator at 74 units of production per hour to the fastest at 153

In 1937 a battery of tests was administered to the bookkeepers but it was not possible at the time to secure a satisfactory criterion In 1941, after the criterion problem had been satisfactorily solved, another battery of 22 additional tests was administered, with excellent results

Description of the Job

The job is the operation of the standard bank bookkeeping adding machine The debits posted are the checks cashed by the customer and the credits are his deposits

A detailed time and motion analysis of the job is contained in Table 3 1, which shows that there are five distinct operations which may be broken up into 18 motions It will be noted that this is a bi manual job This table also refers to the tests used which are listed by name in Table 3 3, some of which it was hoped would show good correlations with the criterion

* Reprinted from *Journal of Applied Psychology,* Vol 27, No 6, December 1943

The Criterion

Considerable attention was devoted to the selection of the criterion, its measurement and its reliability Other investigators have adopted an error score of machine operators as the most valid criterion of performance[1] We have selected instead the speed of posting

The relative value of a bookkeeper depends chiefly on the *amount* of satisfactory work turned out in a given period of time Operators cannot be permitted to remain at work if they make many errors Consequently they learn to work at that speed at which they make few or no mistakes

As a matter of interest, the correlation of production records (shown in Table 3 5) with the corresponding error records is 02, and the correlation between Otis scores and errors is 11 This error score was not, however, made up in the same manner as the one referred to in footnote 1

Table 3 5 is a list of the operators in May, 1941, then or formerly on the machines, arranged in order of their speed of production as averaged from three measures taken in October, 1939, April, 1940, and December, 1940 This is the criterion group It is composed of all girls who were

[1] W H Stead and C C Shartle *Occupational Counseling Techniques* New York American Book Company 1940 See pp 147, 149 for a study of 52 bank bookkeeping machine operators The criterion used was per cent efficiency scores computed from errors, etc, and a multiple correlation coefficient of 45 was obtained with five tests

TABLE 3 1

Operation and Motion Analysis of Machine Bookkeeper

Operation No	*Operation*	*Avg Time*	*Motion No*	*Left Hand*	*Eyes*	*Right Hand*	*Tests Tried*
1	Select ledger card	1 3 sec	1	Rest	Look at cards	Search for card	1 4 5 6 7 8 10 14 16 20 2 23 25
			2	Rest	Look at cards	Find card	
			3	Move to carriage	Look at cards	Grasp card	
			4	Rest	Move to carriage	Carry card	4
2	Insert ledger card	1 1	5	Grasp card	Line up card and guide	Position card	17 20 – –3 5
			6	Insert card	Line up last balance	Insert card	
			/	Return to keys	Read last balance	Move to carriage lever	2 3 9 15 18
			8	Rest	Move to keys	Depress carriage lever	
3	Pick up old balance	1 1	9	Depress keys	Read keys	Depress keys	Same
			10	Move to check	Read check	Depress motor bar	
4	Post check amount	2 2	11	Grasp check	Read keys	Depress keys	
			1	Turn check	Read name next check	Depress motor bar	
			13	Release check	Read name next check	Move to carriage	
			14	Return to keyboard	Read name next check	Rest	
			15	Depress balance key	Read name next check	Rest	
5	Return card	1 1	16	Rest	Move to carriage	Grasp card	
			17	Rest	Turn to tray	Carry card	
			18	Rest	Look for next card	Release card	
		6 8 sec					

able to take the additional tests given in May, 1941, and of whom there were production records at one or more of these three times, and who had been on the machines for at least eight months on the dates production was measured

Production scores follow the normal distribution pattern rather closely, considering the size of the sample

Production Scores	*Frequency*
130–	2
120–129	4
110–119	15
100–109	14
– 99	4
	39
Mean	111 4

Each operator kept a record of the time required to post to ledger and statement cards and a count of the number of b il ances extended and the numbers of debits and credits posted Two operators worked successively with the same items, one posting to the statement sheets and the other to the ledger sheets, so that all figures were verified The raw data were transformed into production units through the use of a formula developed by The Burroughs Company By the use of this formula the standard minutes for the number of operations performed can be calculated When this figure is divided by the number of actual minutes required to perform the operation the production score is derived An operator who works at the "standard" rate will have a production score of 100 The formula reads as follows

$$\frac{\text{Balances} + 6\ (\text{Debits} + \text{Credits})}{30} = \text{Standard Minutes}$$

$$\frac{\text{Standard Minutes}}{\text{Actual Minutes}} = \text{Production Score}$$ '

RELIABILITY OF THE CRITERION

The reliability of each of the production trials was tested by determining the consistency between amounts of work done on the first and third days with the amounts done on the second and fourth days The original and corrected coefficients are shown in the following table

	Trial I Oct 1939	Trial II Apr 1940	Trial III Dec 1940
Number	32	28	31
Range	94–124	94–153	95–126
r obt	88	80	96
r corr	93	85	98

The reliability of the final averages cannot be determined exactly since they are based upon data of varying reliability However, it may be confidently regarded as satisfactory, although the consistency between the various trials is somewhat less than the consistency of each trial taken by itself The inter trial coefficients of correlation are shown in the following table

	N	r	PE
Trial I with Trial II	26	83	± 04
Trial I with Trial III	24	79	± 05
Trial II with Trial III	24	72	± 06

THE TESTS

Beginning in 1936, new bookkeepers were required to score 130 on both parts of the Minnesota Clerical Test and 32 on the 20 minute Otis S A Test, Form B This had little effect before 1941 because of small turnover of bookkeepers In 1941 all bookkeepers were given 22 other tests The correlations between the criterion and the scores on these tests are shown in Table 3 3

All except three of the tests are well known standard tests Test 18, Number Writing, is essentially the subtest in the I E R Clerical, modified by an arrangement of squares into which the numbers must be put in such a way that all numbers are in vertical alignment from right to left Number 5 Name Finding, is a test designed to measure alphabetical selection or filing ability, devised by Mr Paul K Fryer Number 10, Cook s Filing Test,[2] is a drawer containing 3 ′ × 5 ′ index cards, which are pierced with a rod in such a way that when a card is lifted at one end it can be tipped at 90 degrees without being removed from the drawer The drawer contains about 800 cards arranged in alphabetical order The subject is given a list of 100 of these names arranged in random order and the score is the number of correct cards turned up in 10 minutes This test correlated 43 with the criterion but is not being used because there was a significant difference between test scores of experienced and inexperienced operators in favor of the experienced girls, the critical ratio being 5 1 This test is nearly identical with the job and suggests that experience on the job gives practice in the test The advantage which experienced operators have on this test is an example of the danger of constructing a test too much like the job Unless such a test is separately

[2] Loaned by David W Cook of Western Electric Co , Kearney N J

standardized for inexperienced operators it will be of no value in predicting success, but will only discriminate between the better and poorer experienced operators

It will be noted that the form of Army Alpha used was Guilford s Nebraska Revision [3]

able who had been retested on Form B' The total number of cases was 85 and the interval between tests was from eight months to six years The correlation between first and second sets was 86

Alpha Number Series The split half method produced an r of 79, corrected r

TABLE 3 2

Comparison of Test Characteristics for Different Groups

Alpha Number Series	*Range*	*Mean*	*Standard Deviation*
a 39 experienced bookkeepers tested	0–16	9 1	3 6
b 57 bookkeepers tested (experienced and inexperienced)	0–16	9 7	3 3
c 157 women who answered an advertisement, not experienced in bookkeeping	1–17	10 0	3 0
Fryer Name Finding			
a 39 experienced bookkeepers tested	9–22	14 7	2 8
b 57 bookkeepers tested (experienced and inexperienced)	9–22	14 7	2 8
c 157 women who answered an advertisement, not experienced in bookkeeping	4–22	13 5	3 4
d 118 women clerical applicants not experienced in bookkeeping	7–24	14 2	3 6
Minnesota Numbers			
a 39 experienced bookkeepers tested	94–196	147 2	23 8
b 241 experienced women clerical applicants	74–191	126 5	24 2
c 229 inexperienced women clerical applicants	58–188	119 0	23 0
d 1461 experienced and inexperienced women clerical applicants	60–199	128 5	25 5
Minnesota Names			
a 36 experienced bookkeepers	64–186	138 8	22 5
b 241 experienced women clerical applicants	66–190	132 1	25 5
c 229 inexperienced women clerical applicants	52–192	124 1	26 1
d 1472 experienced and inexperienced women clerical applicants	30–199	131 5	27 8

being 88 with N of 57 Four months later a re test was made, with the correlation between first and second scores of 86 and N of 48 Comparison of mean scores between two groups of bookkeepers, in one of which all girls had more than eight months experience, and in the other of which all were inexperienced, showed a significant difference in favor of the inexperienced group When two groups, experienced and inexperienced, were equated on Otis scores, and then compared, the difference between the mean scores on

Reliability of the Tests

A study was made of the reliability of the tests that are used in Batteries I and II, with the following results

Otis S A, Form 'B" (*20 min*) In addition to the 39 cases in the bookkeeping machine operator group, others were avail

[3] J P Guilford, "A New Revision of the Army Alpha Examination and a Weighted Scoring for Three Primary Factors, *Journal of Applied Psychology,* 1938, Vol 22, 239–251

Number Series became insignificant This is explained by the circumstance that the experienced group included most of the slower operators, many of whom made very low Otis and Alpha scores They were employed before tests were used

Fryer Name Finding Split half r was 71 for 68 cases and 83 corrected Re test r was 73 for 65 cases after four months Difference between mean scores of experienced vs inexperienced girls was not significant

Minnesota Number Checking Re tests were available for 36 of the 39 bookkeeping machine operators and produced an r of 61 ± 06 Of these 36 cases, 14 showed an increase in test scores, 16 a decrease and 6 were unchanged (to within a 3 point difference) A group of 77 women clerks who were first tested at the time of employment were re tested 8 to 16 months later The re test r was 69 ± 04 Another similar group of 59 women clerks, re tested after 17 to 54 months, or an average of 38½ months, produced a re test r of 56 ± 06

Minnesota Name Checking Comparison of test and re test scores for 36 machine bookkeepers gave a correlation of 75 ± 05

The group of 77 new women employees produced a re test r on the name test of 62 ± 05 The group of 59 produced an r of 81 ± 03

The statistics of all three groups on number and name tests are as follows

tested during their summer Saturday holidays and they were paid for their time at the rate of $1 00 per hour This solved the problem of finding the time for testing and produced very good cooperation from the girls

RESULTS

Table 3 3 shows the correlations between the criterion and all tests It will be noted that all the higher coefficients of correlation are of tests of verbal or numerical material, or tests with these two kinds of material mixed

Six of the tests were of hand, arm and finger dexterity The operation analysis of the job revealed a good many such movements and it appeared likely that one of these tests would show considerable relationship with the criterion The fact that none did suggests that the mental abilities required are so overwhelmingly important that the dexterities can be ignored Marion A Bills, in correspondence with the author, reports similar findings

Intercorrelations were calculated for all tests The battery that had been in use for several years, namely the Otis and the two Minnesota tests, produced a multiple R of + 65 ± 06 Good results were also obtained with a battery consisting of Alpha Number Series, Fryer Name Finding and Minnesota Numbers This gave an R of + 70 ± 05 Table 3 4 shows the intercorrelations of the tests that went into the

	36 bookkeepers		*77 women clerks*		*59 women clerks*	
	Numbers	Names	Numbers	Names	Numbers	Names
r_{12}	61 ± 06	75 ± 05	69 ± 04	62 ± 05	56 ± 06	81 ± 03
M_1	146 9	138 8	137 1	142 3	146 3	142 8
σ_1	22 2	22 5	19 3	16 5	16 1	25 3
Range$_1$	94–192	64–186	96–196	108–192	80–198	64–194
M_2	145 1	133 9	151 1	155 8	152 2	144 3
σ_2	22 9	25 1	20 9	18 6	18 1	28 7
Range$_2$	92–194	62–170	98–196	113–194	92–194	62–194

ADMINISTERING THE TESTS

The time required to give the complete battery of tests was about four hours and it was a problem to take the bookkeepers away from their work for this long Finally, arrangements were made to have the girls

Wherry Doolittle multiple correlation formula, and gives the range, standard deviation and mean of the five tests used in the two multiple batteries just mentioned These two batteries are now being used in the selection of new bookkeepers

TABLE 3 3

Correlations for 39 Bookkeepers

Test	Correlation With Production	Probable Error of Correlation
1 Otis S A 20 min	+ 56	± 07
2 Alpha Number Series	+ 56	± 07
3 Minnesota Numbers	+ 51	± 08
4 Alpha Total	+ 51	± 08
5 Fryer Name Finding	+ 48	± 08
6 Alpha Same—Opposites	+ 47	± 08
7 Minnesota Names	+ 47	± 08
8 Alpha Verbal	+ 47	± 08
9 Alpha Numerical	+ 44	± 09
10 Cook s Filing	+ 43	± 09
11 Alpha Relationships	+ 43	± 09
12 Alpha Analogies	+ 42	± 09
13 Alpha Information	+ 40	± 09
14 Alpha Sentences	+ 40	± 09
15 Alpha Arithmetic	+ 37	± 09
16 Alpha Directions	+ 32	± 10
17 Worker Analysis Bi Manual Peg Board	+ 24	± 10
18 Fryer Number Writing	+ 22	± 10
19 Alpha Judgment	+ 20	± 10
20 O Connor Pinboard	− 09	± 11
21 Ziegler Rate of Manipulation, Left	− 06	± 11
22 Worker Analysis Washer—Assembled	− 05	± 11
23 Ziegler Rate of Manipulation Turning	− 04	± 11
24 Ziegler Rate of Manipulation Right	− 03	± 11
25 Worker Analysis Washer—Unassembled	− 01	± 11
Years of Experience	+ 05	± 11

TABLE 3 4

Intercorrelation for 39 Bookkeepers

	Otis	Alpha Arith-metic	Alpha Same Opposite	Alpha Sen tences	Alpha Number Series	Alpha Anal ogies	Fryer Name Finding	Minne sota Numbers	Minne sota Names
Production	56	37	47	40	56	4	48	51	47
Otis		54	71	72	68	78	55	36	41
Alpha Arithmetic			53	4	68	38	41		17
Alpha Same Opposite				71	59	54	5	[illegible]	15
Alpha Sentences					55	6 1	44	17	21
Alpha Number Series						67	35	34	56
Alpha Analogies							34	0	36
Fryer Name Finding								3	38
Minnesota Numbers									[illegible]
Minnesota Names									
Range	17–59				0–16		9–22	94–196	64–186
Standard Deviation	10 3				3 6		2 8	-3 8	-4 5
Mean	32 5				9 1		14 7	147	138 7
P E meas	- 6				0 9		1 0	10 1	8 (
Experience									
Range	9 months to 19 years								
Mean	9 2 years								

	Battery	*R*
I	Alpha Number Series, Fryer Name Finding, Minnesota Numbers	70± 05
II	Otis 20 min , Form B, Minnesota Numbers, Minnesota Names	65± 06
III	Otis 20 min Form B, Minnesota Numbers, Alpha Number Series, Fryer Name Finding	71± 05
IV	Otis 20 min , Form B, Minnesota Names	62± 07
V	Minnesota Names, Alpha Analogies, Alpha Sentences	56± 07
VI	Otis 20 min Form B, Minnesota Names, Alpha Sentences	62± 07
VII	Otis 20 min , Form B, Minnesota Names, Alpha Arithmetic	62± 07
VIII	Otis 20 min , Form B Alpha (total score) Minnesota Names, Minnesota Numbers, Fryer Name Finding	68± 06

Multiple correlations were calculated for several other combinations Battery III was the result of the use of the Wherry Doolittle method into which calculations went the scores of all tests shown in Table 3 4 All other batteries were calculated by the Doolittle method

A practical expression of the efficiency of Battery I may be seen by reference to Table 3 5, which lists the criterion, or ac tual production records, in contrast with the predicted production resulting from the use of Batteries I and II Battery I pre dicts success for 19 of 21 operators whose production record is 110 or better, a per centage of 91 Battery II predicts 81 per cent of the successful operators Both bat teries predicted success for 5 of the poorer operators, a 'miss of 28 per cent Battery III, resulting from the use of the Wherry Doolittle method, was successful in pre dicting 85 per cent of the successes and missed on 28 per cent of the failures

One critic of this paper has suggested that the correlations that were obtained would have been higher if the range of performance of the group had not been somewhat small due, no doubt, to a homo geneity of the group resulting from long service on the job The correlations ob tained are probably 'lower than if an unselected group were hired and given equal chances to succeed or fail

Another critic refers to another experi ment in which a test gave a positive cor- relation with success when applied to those who were already on the job, but subse quently gave none when applied to newly hired persons He therefore suggests caution in investigating differences in scores be tween our experienced group and newly hired inexperienced persons Fortunately, such comparisons are available and, except for tests which were discarded, no signifi cant differences were discovered that seem to relate to our use of the tests

Correlations were calculated with the Powers tabulating machine Despite the lack of previous experience in running correlations with this equipment, it seemed advisable to give it a try in the expectation that the work of calculating intercorrela tions for 25 tests would be less than if scattergrams or calculating machines were used It transpired that much more labor was involved than was anticipated, the Powers equipment not being as well adapted to the work as is the Hollerith machine Calculations of *r* s were from un- grouped scores

The prediction formula for Battery I, corrected for dispersion, is

$$\overline{X}_0 = 1\ 34 \times \text{No Series} + \ 19 \times \text{Minn Nos} + 1\ 27 \times \text{Name Finding} + 52\ 43$$

and for Battery II is

$$\overline{X}_0 = \ 59 \times \text{Otis} + \ 19 \times \text{Minn Nos} + \ 05 \times \text{Minn Names} + 57\ 84$$

TABLE 3 5

Comparison of Actual and Predicted Production

Operator Number	Actual Production Record	Production Predicted by Battery I (Number Series, Minnesota Numbers, Name Finding)	Production Predicted by Battery II (Otis, Minnesota Numbers, Minnesota Names)
1	140	131	129
2	130	122	122
3	129	118	121
4	124	117	114
5	123	112	108
6	120	128	130
7	119	111	113
8	118	114	121
9	118	122	120
10	117	108	104
11	117	121	117
12	115	97	99
13	115	117	119
14	114	119	126
15	113	110	101
16	113	119	122
17	113	114	119
18	113	125	116
19	112	126	117
20	110	112	123
21	110	113	116
22	108	108	109
23	108	110	114
24	108	112	108
25	106	108	100
26	106	102	103
27	106	102	101
28	106	106	102
29	105	107	114
30	104	105	100
31	104	111	109
32	104	119	118
33	103	105	104
34	103	100	103
35	103	104	99
36	98	110	113
37	98	89	94
38	96	103	100
39	94	87	96
Mean	111 4		

ADDITIONAL DATA

A steady improvement in the average production of bookkeepers has taken place since tests were first used The following table shows the average production record of all operators on whom records were available on the dates shown The lower number of operators at the later dates is due to the presence of new and unseasoned workers

Date	*N*	*Average Production*
October 1937	43	105 0
November 1939	40	108 7
April 1940	30	109 7
December 1940	32	110 2
December 1941	26	110 9

SUMMARY

The problem was to find tests from whose scores the future success or failure of bookkeeper job applicants could be predicted A reliable criterion was obtained and 25 tests were administered to bookkeepers then on the job By the use of multiple correlation several efficient test batteries were identified, the best ones of which predicted 91 per cent of the better operators and 72 per cent of the poorer ones Two of these batteries have been used for nearly five years with excellent results

Postscript to Predicting Success in Machine Bookkeeping *

EDWARD N HAY

In December 1943, a successful experiment in validating tests for predicting success in machine bookkeeping for a large commercial bank was reported [1] Since 1941, when that experiment was completed, repeated measures have been made of the performance of individual bookkeepers The data show that there has been a substantial improvement in production year by year, as a result of the employment of operators selected by test Table 4 1 gives the detailed record chronologically

TABLE 4 1

Production Rates of Machine Bookkeepers
Burroughs Index of Production Rate

Date	*N*	*Average Production*
October 1937	43	105 0
November 1939	40	108 7
April 1940	30	109 7
December 1940	32	110 2
December 1941	26	110 9
December 1943	29	116 7
June 1944	35	113 0
December 1945	29	108 1
June 1946	29	111 6

N in each case is the number of bookkeepers who have been on the machine at that time for one year or longer The decline in the performance from a high of 116 7 in 1943 to a low of 108 1 in December 1945 was the result of a smaller supply of applicants with acceptable test scores The subsequent increase to June 1946 of 111 6 reflects the improvement in the supply of higher test score operators since the end of the war

* Reprinted from *Journal of Applied Psychology* Vol 31, No 3, June 1947

[1] Edward N Hay Predicting Success in Machine Bookkeeping' *Journal of Applied Psychology* 1943, Vol 27, 483–493

Relation Between Scores on Certain Standard Tests and Supervisory Success in an Aircraft Factory *

A Q SARTAIN

The writer wishes to express his appreciation to the Texas Division of North American Aviation, Inc , for supporting and making possible this study Special acknowledgment is made of the help of Mr Ross A Peterson, Director of Education

The question of how to select supervisory personnel is frequently one of the most important faced by a business enterprise Since success as a worker is no guarantee of success in supervision, it is natural that psychological tests should be considered as possible instruments for selection of suitable persons for supervisory responsibilities

Statement of the Problem

The problem of this study was to determine the extent to which success in supervision in an aircraft factory was predicted by the following standard tests Otis Self Administering Test of Mental Ability (Higher Examination), Tiffin and Lawshe Adaptability Test (Form A), Revised Minnesota Paper Form Board, Bennett Test of Mechanical Comprehension (Form AA), Remmers and File How Supervise? (Experimental Edition, Form A), Bernreuter Personality Inventory, and Kuder Preference Record

Subjects and Conditions of the Experiment

The tests listed above were given to 40 members of supervision in the factory Thirty seven of these men were assistant foremen, and three were foremen Each man was rated by the foreman and general foreman over him (except in the case of the foremen, where it was necessary to secure a second rating by the general foreman, the second rating being obtained about three weeks after the first) Each man was rated on two different rating forms, and the combination of the four ratings constituted the criterion of success

The Criterion

In setting up the criterion, the ratings on each rating form were converted to standard deviation scores, and the sum of these scores became the criterion An attempt was made in preliminary studies to insure both the reliability and the validity of each rating form One of these forms (called Form A henceforth) consisted of the seven qualities which had been found to correlate most highly with success as a supervisor, each quality being listed on a separate sheet In the preliminary study, the correlation between the average of two ratings and the average of two scores or grades given for success on the job was 88 (84 when new ratings were secured five weeks later), and the correlation between two ratings for each man was 64 The number of employees involved was 43 Thus, it is concluded that Form A was sufficiently reliable and valid to comprise a part of the criterion

The second rating form (Form B) consisted of ten qualities, all on a single sheet In the preliminary study ($N=54$), the correlation between the average of two ratings and the average of two scores or grades on supervisory success was 92 The two ratings correlated with each other to the extent of 63 Thus, it appears that Form B was also reasonably reliable and valid

It should be emphasized that the results just cited were from earlier studies of the rating forms In the present study the results were hardly so favorable Table 5 1

* Reprinted from *Journal of Applied Psychology* Vol 30, No 4, August 1946

TABLE 5 1

Correlations between Ratings Constituting Criterion

Ratings	r
Average Rating on *A* vs Average on *B*	79
First Rating on *A* vs First on *B*	77
Second Rating on *A* vs Second on *B*	62
First Rating on *A* vs Second on *B*	54
Second Rating on *A* vs First on *B*	48

presents the relevant findings for this study While these correlations are not as high as earlier studies might lead one to expect, they appear to be high enough to indicate that the combined ratings might well serve as the criterion

Results of the Study

As Table 5 2 brings out, correlations be tween the test scores and the criterion were low, in every case so low as to lack sta tistical significance (According to Fisher, for a coefficient of correlation to be signifi cant at the 5 per cent level of confidence under the conditions of this study it would have to be 304, at the 1 per cent level of confidence it would have to be 393 [1]) These low correlations may be due to a faulty criterion It seems more probable, however, that the tests simply fail to cor relate with supervisory success in this plant

Correlations were obtained between some of the test scores, and are presented in Table 5 3 The correlation between the two general mental ability tests (86) and those between the mechanical ability tests and the general mental ability tests (33 to 41), as well as that between the two mechanical ability tests (31), are not far different from those found in most similar studies [2] The correlation between Adaptability and How Supervise? indicates that general mental ability goes with favorable super visory attitudes (low scores on this test in dicating a favorable attitude) to a moder ate degree Other coefficients are too small to have significance

TABLE 5 2

Coefficients of Correlation between Test Scores and Criterion

Test	r
Otis Self Administering	04
Adaptability	− 07
Minn Paper Form Board	10
Bennett Mechanical Comprehension	− 15
How Supervise?	− 18
Bernreuter Personality Inventory	
B1 N	− 11
B4 D	12
F1 C	01
F2 S	07
Kuder Preference Record*	
Mechanical	004
Social Service	− 06
Clerical	003

* The plant was closed before this study was concluded and the data on the other interest scales of the Kuder test inadvertently destroyed

[1] J P Guilford *Psychometric Methods* New York McGraw Hill Book Co, Inc, 1936, p 549

[2] E B Greene, *Measurements of Human Behavior* New York Odyssey Press, 1940 pp 257, 361

TABLE 5 3

Coefficients of Correlation between Certain Test Scores

Tests	r
Adaptability vs Otis	86
vs How Supervise?	− 44
‘ vs Form Board	33
‘ vs Bennett	41
Otis vs Form Board	39
‘ vs Bennett	37
Bennett vs Form Board	31
How Supervise? vs Kuder Persuasive	00
‘ vs Kuder Social Service	17
Kuder Mechanical vs Form Board	13
‘ vs Bennett	15
Kuder Scientific vs Form Board	19
“ ‘ vs Bennett	15

ADDITIONAL STUDIES

Two other studies were made of the success of the Otis and Bernreuter in selecting supervisors In one of these, the sum of the scores on both the rating scales was again used as the criterion of success, two ratings on each form being secured on 85 men Table 5 4 is based on this study It is clear that the coefficients are most likely due to chance

In the second study, 53 members of supervision who were known well to three

TABLE 5 4

Relation of Bernreuter and Otis Scores to Rated Success in Supervision

Test	r
Otis	16
Bernreuter	
B1 N	− 12
B4 D	04
F1 C	− 09
F2 S	− 02

TABLE 5 5

Comparison of Bernreuter and Otis Scores of Groups of Good and Poor Supervisors

Test or Scale	Poor Group				Good Group				Critical Ratio
	No	Mean	S D	$S D_M$	No	Mean	S D	$S D_M$	
B1 N	29	−127 1	61 20	11 56	24	−146 4	46 30	9 66	1 29
B4 D	29	85 6	48 50	9 16	24	108 3	65 10	13 58	1 39
F1 C	29	− 95 0	73 05	13 77	24	−109 3	51 90	10 80	82
F2 S	29	− 36 5	44 20	8 35	24	− 38 8	48 90	10 18	17
Otis	28	101 1	13 03	2 51	24	105 1	10 08	2 10	1 22

individuals in management positions were divided into two groups, good supervisors ($N = 29$) and poor supervisors ($N = 24$) The members of each group were selected because there was agreement among those classifying them that they belonged in one or the other group When the Bernreuter and Otis scores of these two groups were compared, the results shown in Table 5 5 were obtained It will be noted that the differences all favor the good supervisors, that is, that they appear to be more intelligent more stable, more dominant, more self confident, and more sociable, but that no difference even approaches statistical significance

SUMMARY AND CONCLUSIONS

The following tests were administered to forty members of supervision in an aircraft factory Otis Self Administering Test of Mental Ability (Higher Examination), Tiffin and Lawshe Adaptability Test (Form A), Revised Minnesota Paper Form Board, Bennett Test of Mechanical Comprehension (Form AA), Remmers and File How Supervise? Test (Experimental Edition, Form A), Bernreuter Personality Inventory, and Kuder Preference Record Two ratings on each of two rating forms were then secured for each man, the rating forms previously having been checked for reliability and validity and the sum of the four ratings (reduced to standard deviation scores) became the criterion of success Test scores were then correlated against the criterion In every instance the coefficients obtained were too low to be considered significant, the highest one being only 18 It was concluded, therefore, that these tests had little or no predictive value for success in supervision in this plant

Two additional minor studies corroborating this conclusion in part are also reported

Testing Programs Draw Better Applicants *

ELEROY L STROMBERG

SUMMARY

A number of personnel managers who have adopted selective test procedures have been surprised and even alarmed to discover that within a short time almost all applicants qualify on the test batteries, even though the installations were carefully made and flexible selection standards established No great problem arises for the employment manager can effectively adjust his critical scores to meet the fluctuation of the labor market, however, the reasons for the superiority in the test level of new applicants as compared to the original criterion groups present an interesting problem to be considered Although the answers are not clear cut, the implications for such changes in qualifications are the subject of this report These indicate that the existence of a testing program attracts the better applicants and discourages the poorer

A selective testing program was installed for all productive workers in three plants of the same industry One plant was in Illinois, one in Maryland, and the third in Washington, D C All three plants are members of the same national association This is a borderline industry, employing women primarily, and data reported are for women only The wage scale is low The Illinois plant employs only white women while about 95 per cent of the productive workers in the east coast plants are Negro women

Selection tests were validated by testing

* Reprinted from *Personnel Psychology,* Vol 1, No 1 spring 1948 This paper was read at the Industrial section of the Midwestern Psychological Association in May 1947 (1, 2)

present employees in all three plants Subsequent applicants were superior to the criterion groups to such a degree that initial standards failed to screen out the expected number of applicants The differences between the experimental groups and subsequent applicant groups are statistically significant These differences cannot be explained in terms of shifts in the labor market or motivation It is suggested that self imposed selection may take place within the labor source, and that only the more capable applicants apply at plants where it has become known that test qualifications have been made a requisite for employment

The Selection Tests

The study was begun in the Illinois plant in the fall of 1945 Several tests which might be useful in the selection of job applicants were selected or constructed The productive workers in this plant were rated on a graphic rating scale (especially prepared for the study) by a supervisor and by the general manager of the plant The author served as recorder during the rating procedure Ratings were secured for approximately 130 workers who had been employed for periods long enough to make rating possible Criterion groups comprising the upper and lower 25 per cent of the productive workers were selected on the basis of these ratings The measuring devices were given their initial trial with these criterion groups Three tests proved to be discriminative, and it was decided to use them in the employment office One of these was the Purdue Adaptability Test (5), the second was the Code Identification Test, which is a test of ability for a specific operation in this industry, and the third was a test of personal adjustment composed of a number of items similar to the less extreme items of the Minnesota Multiphasic A scoring key was prepared for the personal adjustment test on the basis of an item analysis of the responses made by the workers in the two criterion groups

Comparing Applicants and Employees

Beginning in February, 1946, all applicants were given the three tests The data in Table 6 1 summarize the results (in the Illinois plant) for the criterion group and for applicants during the six month period from February through July, 1946 For the Adaptability Test and the Code Identification Test the mean score for the applicant group is higher than for the criterion group On the test of adjustment, in which a low score is more desirable than a high score, the applicant group is also superior The probability that such differences in favor of the applicant group could arise by chance is extremely low

The second plant, located in Maryland, employs more than twice as many productive workers as the Illinois plant Criterion groups were selected in the same manner as in the first plant The author served as recorder while the department foreman and the floor superintendent indicated their ratings of the workers on the graphic scale An additional test, the Discriminative Dex

TABLE 6 1

Comparison of a Criterion Group and the Applicants in the Subsequent 6 Month Period

		Code Identification	*Purdue Adaptability*	*Personal Adjustment*
Illinois Criterion Group	M	27 87	8 89	10 45
	σ_M	2 36	69	1 41
	N	54	56	44
Illinois Applicant Group	M	35 47	11 19	8 42
	σ_M	1 88	52	50
	N	71	71	65
	P	02	01	02

terity Test (4), was introduced in this plant This is a test of dexterity involving somewhat the same principle as the Minnesota Rate of Manipulation Test but differing in that no two successive movements made by the examinee are alike

The data for the criterion group and for the applicant group of the Maryland plant are presented in Table 6 2 The criterion group represents an employed sampling in March, 1946 The applicant group represents those applicants tested (and all were tested) from April through June, 1946 The mean scores on the Code Identification Test and the Adaptability Test are higher for the applicant group than for the criterion group Both on the Adjustment Test and on the Dexterity Test, the applicant group is again superior to the criterion group The probability of chance differences is very low except for the Adjustment Test

In March, 1946, when the Maryland selective battery had just been completed, newspaper advertising attracted approximately 50 applicants who were tested on the four tests Their scores were compared with those of the criterion group in order to determine whether or not the local labor market would justify the establishing of certain critical selection scores

The data for the small group of control applicants in March and for the criterion group measured in the same month are shown in Table 6 3 The control applicants made poorer scores on the Code Identification Test and on the Adaptability Test Both of the differences in favor of the criterion group are significantly greater than would be expected to occur as a result of sampling factors alone On the adjustment test and the Discriminative Dexterity Test the control applicant group made scores superior to those of the criterion

TABLE 6 2

Comparison of a Second Criterion Group and the Applicants in the Subsequent 3 Month Period

		Code Identification	*Purdue Adaptability*	*Personal Adjustment*	*Discriminative Dexterity*
Maryland Criterion Group	M	22 58	5 64	9 07	198 70
	σ_M	1 38	34	1 60	3 47
	N	100	101	78	100
Maryland Applicant Group	M	30 83	7 08	6 05	163 36
	σ_M	92	31	40	2 15
	N	205	204	213	164
	P	01	01	07	01

TABLE 6 3

Comparison of a Criterion Group and the Applicants Responding to Newspaper Advertising During the Test Installation Period

		Code Identification	*Purdue Adaptability*	*Personal Adjustment*	*Discriminative Dexterity*
Maryland Criterion Group (Table 2)	M	22 58	5 64	9 07	198 70
	σ_M	1 38	34	1 60	3 47
	N	100	101	78	100
Maryland Applicant Control Group	M	18 17	3 67	5 28	188 91
	σ_M	1 56	47	1 43	3 42
	N	47	48	32	46
	P	04	01	07	05

TABLE 6 4

Comparison of Immediate (Control) Applicants and the Applicants in the Subsequent 3 Month Period

		Code Identification	*Purdue Adaptability*	*Personal Adjustment*	*Discriminative Dexterity*
Maryland Applicant Control Group (Table 3)	M	18 17	3 67	5 28	188 91
	σ_M	1 56	47	1 43	3 42
	N	47	48	32	46
Maryland Applicant Group (Table 2)	M	30 83	7 08	6 05	163 36
	σ_M	92	31	40	2 15
	N	205	204	213	164
	P	01	01	61	01

group The difference found on the dexterity test in favor of the applicant group is significantly different from chance

A comparison of the control applicants and applicants for the succeeding three months is made in Table 6 4 In this comparison it will be observed that the later applicants are superior to the criterion groups The values indicating the probability of chance differences for the Code Identification Test, the Adaptability Test, and the Discriminative Dexterity Test are identical to those given in Table 6 2 The P value for the Adjustment Test again does not indicate a significant difference between the criterion and applicant groups

WHAT THE RESULTS IMPLY

These results appear to be a warning to those of us who are interested in industrial selection problems If applicant groups subsequent to the installation of a selection procedure are better qualified for the work within the industry than the criterion group of present workers, and superior to a control group of applicants, the often reported increase in efficiency of the new workers cannot be wholly attributed to the selectivity of the tests themselves The tests employed in this study had been validated within the industry in which they were to be used, and it can be assumed that they were able to increase the productivity of the industry through their selectivity However, not only the employees selected but the applicant group as a whole show increases in measured ability following inauguration of the testing program It is possible to assume that any testing program, regardless of its validity could have done as well as those which had been carefully validated through many weeks of effort

The question arises as to why the applicant groups were superior to the criterion groups both among white women workers in Illinois and Negro women workers in Maryland Ordinarily the changes in the labor market are not so rapid that such large changes should occur within a period as short as three months A check of the labor market was made by comparing the criterion group in the Maryland plant with a criterion group in the Washington, D C, plant of the same company In the Washington plant only Negro women are employed Data for the criterion group in Washington were gathered in July, 1946, and are compared to the Maryland criterion sample in Table 6 5 Only in the case of the Discriminative Dexterity Test is the difference greater than might occur through chance factors in the samples

Perhaps the applicants are more highly motivated than the workers already on the job This may be the reason for the reduced time scores on the dexterity test but it can hardly account for the higher scores made by the applicants on the Adaptability Test and Adjustment Test Moreover, the control applicant group tested at the same time as the criterion group in Maryland (Table 6 3) does not show evidence of this added motivation

TABLE 6 5

Comparison of a Criterion Group in March and a Criterion Group in a Neighboring City 4 Months Later

		Code Identification	*Purdue Adaptability*	*Personal Adjustment*	*Discriminative Dexterity*
Maryland Criterion Group (Table 2)	M	22 58	5 64	9 07	198 70
	σ_M	1 38	34	1 60	3 47
	N	100	101	78	100
Washington D C Criterion Group	M	23 28	4 68	9 66	183 21
	σ_M	1 52	38	73	2 56
	N	110	114	1 02	112
	P	74	06	80	01

One explanation appears more reasonable than the others and that is that each of these industries tends to draw its applicants from a specific labor group, and perhaps from a specific local population area The distributions of the test scores for the applicants are conspicuous by virtue of the absence of the expected frequency of very low scores on any of the tests Perhaps those individuals who ordinarily have difficulty in the test situation do not now apply The information that You have to take a test to get a job there' may keep the less desirable individuals from applying Likewise this may be an incentive which attracts some who would not otherwise be interested in applying at a borderline industry

Even if these possible explanations for the increased qualifications of applicants are correct, there is a more important problem to be faced by industrial psychologists If the inference that any test might have been as effective in increasing the quality of the applicants is true, then the industrialist may become more cautious in the expenditure of his money for 'tailor made' programs, or for the validation of specific tests for his particular selection problem[1] This suggests that psychologists should place increasing emphasis on the placement function of psychological tests A selection program that is concerned only with the elimination of the poor risk fails to make full use of the test data The proper placement of those with special qualifications is equally important if testing programs are to make their maximum contribution

Since this report was presented at the Midwestern Psychological Association meeting an attempt has been made by Rothe (3) to account for similar increases in applicant test scores in terms of motivation On the basis of a comparison between mean scores of criterion groups and later applicants which show the same superiority of applicants as reported in the present study, he rules out all explanations of these changes save one on logical' bases The only explanation offered by Rothe is that applicant groups are more highly motivated, or to use his term incentivated,' than the criterion groups No data to substantiate this opinion are presented The data in Table 6 3 above, from the applicant group tested before information concerning the use of selective tests in the industry had been circulated, contradicts Rothe s assumption The control applicant group is significantly different from later applicants (Table 6 4)

While the author is in complete agreement concerning the necessity for continuing validation of test installations, the inferences of the present report still stand If, whether by means of added motivation, which seems unlikely in view of the data presented above, or by self imposed selection, more adequate applicants seek employment after tests are installed, then any

[1] JOURNAL EDITOR'S NOTE Selection *among* applicants, however is useful in improving worker efficiency only if the battery is one which actually predicts worker success

battery of tests might prove an asset to the industry [2]

REFERENCES

1 Buxton, C E, The Nineteenth Annual Meeting of the Midwestern Psychological Association, *American Journal of Psychology* 1947, Vol 60, 440–442
2 —— Proceedings of the Nineteenth Annual Meeting of the Midwestern Psychological Association, *American Psychologist* 1947, Vol 2, 419–428
3 Rothe, H F, Distributions of Test Scores of Industrial Employees and Applicants, *Journal of Applied Psychology* 1947, Vol 31, 480–483
4 Stromberg E L, Manual for Use With the Discriminative Dexterity Test (Personnel Research Institute Western Reserve University, Cleveland, 1946)
5 Tiffin, J, and Lawshe, C H, The Adaptability Test A Fifteen Minute Mental Alertness Test for Use in Personnel Allocation *Journal of Applied Psychology*, 1943, Vol 27, 152–165

[2] See NOTE 1, page 31

Use of the "Group Situation Observation" Method in the Selection of Trainee Executives *

RONALD TAFT

A recurrent problem in the planned programs for the selection and training of young executives is that of predicting the likely future development of the potential trainee while he is still a youth This article describes the application of the group situation observation technique to this problem of selection This technique was originally used in the German Army Selection procedures, and adopted (and adapted) by the British (3) and Australian Armies (4) The U S Army O S S also utilized the basic principles in connection with the selection of personnel for operations behind enemy lines (6) Since the conclusion of the War, it has been applied to the selection of trainee industrial foremen, managers and civil service administrators, mainly in Britain (1, 2) The present report deals with the application of the technique to a group whose age is well below that of other reported uses (17 to 19 years)

The position for which the candidates were being considered was that of trainee production executive, in a shoe factory with 200 employees Two trainees were required Because of the long-range nature of the training program no exact definition of the traits required by these trainees was attempted, but the selectors were familiar with the factory and the approximate duties which would be required of the future executives

PROCEDURE

The screening procedure prior to the group observation sessions is given briefly to provide a background to the data available to the selectors

Written applications from 63 persons were received as a result of newspaper advertisements, and 13 of these were rejected without interview on educational grounds The Managing Director of the Company then gave an orientation and screening interview to the remaining candidates, as a result of which 11 were rejected as "unsuitable types" and 5 withdrew their applications Eleven failed to report for this interview

The remaining 23 applicants were then given a vocational guidance interview by the writer, at which time they were given the following tests Vocational Interest Questionnaire, Personal Questionnaire 'I' (Hanawalt and Richardson), Oral and Written Directions (Adaptation of Army

* Reprinted from *Journal of Applied Psychology* Vol 32, No 6, December 1948

Alpha), H Test (short form) (Adaptation of Army Alpha), Speed and Accuracy (Minnesota Vocational Test for Clerical Workers), Space Form Perception (Australian Institute of Industrial Psychology), and Mechanical Comprehension (Bennett A A) This was followed by a half hour interview Two failed, however, to report for this interview

Seven more applicants were rejected at this stage on grounds of interest, temperament or ability, including all those with a score of less than the 60th centile on general population norms for the H test

Group Situation Examination

The remaining 14 candidates were invited by mail to be present at the home of the Managing Director to spend the day with him in connection with your application for employment' One failed to attend The others were divided into three separate groups, of four or five, each group being arranged for either a Saturday or a Sunday During the group situation they were under the observation of the Managing Director and the writer (henceforth referred to as the Psychologist), the latter controlling the day s proceedings

The following program was observed

Step	*Time Period*	*Activity*
1	11 45 to 12	Introduction
2	12 to 12 15	Personal History
3	12 15 to 12 45	Game— Who am I?'
4	12 45 to 1 45	Lunch
5	1 45 to 3 30	Group Rorschach Test
6	3 30 to 4 15	Leaderless Discussion
7	4 15 to 4 30	Afternoon Tea
8	4 30 to 5	Problem Situation Discussion
9	5 to 5 30	Personality Judgments of Self and other Candidates
10	5 30	Closing Address by the Managing Director

1 *Introduction* Candidates were welcomed and introduced to each other by the Managing Director, and a brief word on the procedure was given by the Psychologist They were asked to try to adopt an informal attitude and to refer to each other by their first names They were warned that it is 'impossible to beat the system," so that it would be in their best interests to try to be natural right from the beginning rather than to bluff their way through

2 *Personal History* Each candidate was asked in turn to introduce yourself to the others by stating briefly your personal history No further instruction was given, and they were called on in order of age, starting from the oldest At the conclusion of these short outlines, the candidates were given an opportunity of asking questions about the others, but a total of only four questions was asked

This procedure was of some value in giving the candidates a brief outline of their colleagues background, and also in indicating which factors the candidates considered significant in their lives However, there was a tendency to adopt the pattern followed by the first speaker, and it was necessary in evaluating the contributions to consider this factor Thus credit was given to a fourth speaker who broke away from an unsatisfactory habit adopted by the prior speakers of speaking about their schools rather than themselves [1]

Indications about the candidates obtained from this procedure mainly related to self confidence, particularly while in a situation calculated to unsettle them, also their ability to select salient factors, and to follow an independent line

3 *Game—'Who am I?"* Candidates were then informed that they were to play a game commonly known as 'Who am I?,' or 'Personalities They were instructed as

[1] In referring to information obtained as a result of the various tests, the writer has in mind the notes made by the observers at the time, but no attempt was made to infer particular characteristics from the one test only

follows 'One person is to leave the room, and the others are to imagine that they represent a well known personality, either living or dead The person leaving the room should be brought back and should endeavor to find out who the personality is, by asking each one of the others in turn a question the answer to which is either Yes or No' You should keep on asking questions until you have narrowed down the field, and you are allowed only one guess I will not give you any further instructions, and you should work out any other details yourselves Continue with this game until each one of you has had a turn Whenever any questions are asked they were reminded that they were on their own'

This test appeared to be particularly useful as a means of introducing the group to the leaderless group situation, as on each occasion problems regarding the observance of the rules were raised Information obtained from this session related to the ability of the candidates to get their opinions accepted, attitudes towards the observance of rules, flexibility, intelligence, concentration, reaction to frustration, impulsiveness (tendency to guess rather than analyse), persistence, sympathy with the difficulties met by others, extent of general knowledge, identification with famous people, and so on For example when the questioner made an incorrect guess, it was useful to observe how the others responded to the rule that only one guess should be permitted

4 *Lunch* During lunch the Managing Director and the Psychologist endeavoured to take part in the conversation and to make the atmosphere informal Lunch commenced as a standing buffet to permit candidates who were attracted to each other to come together, and a 'mental note was made of their social and individual behaviour

5 *Group Rorschach* This was not properly a group situation test, but was introduced at this stage of the selection procedure for convenience only Also it was felt that doing this test would help to break down tension, by strengthening the feeling on the part of all the candidates that they were going through the same trial together

The use made of the Rorschach interpretations was similar to the use made of the aptitude tests, that is it was primarily a screening device, intended to cull out those with definite neurotic symptoms In this respect one was rejected as too uncontrolled and one as too inhibited, the latter giving only eight responses

6 *Leaderless Discussion* The candidates were seated in a circle, with the Managing Director and Psychologist at the side They were told to regard the latter as merely pieces of furniture," and that they were now to discuss any topic on which they might decide No further instructions were given

The group dynamics involved in the selection of the topic itself provided valuable material This test was also useful for observing how the subject stands up to argument, whether he perseveres or shows resistance to persuasion and whether he becomes emotional In two of the three groups a dominant person seemed to arise at this juncture and an opportunity was afforded for observing whether the form of domination was "autocratic' or 'integrative (in the sense used by Lewin)

7 *Problem Situation Discussion* This discussion differed from the previous one only in so far as it was more structured, that is, the group was given an actual assignment The candidates were given the facts about the hours of work at the factory at which they had applied for the position, and were asked to report their recommendations back to the Managing Director on how they considered these hours should be altered to arrange a 40 hour week (The factory was previously working a 44 hour week)

This discussion again gave scope for observing tendencies in certain of the candidates to dominate their group It also was revealing about the knowledge of the candidates as to the general situation in industry, and their attitude towards management and employees (this was considered in conjunction with their previous experience and home background)

The main difference between the leader-

less discussion and the problem situation discussion is that the former gives more scope for the individual to show his personality and ability *qua* individual, while the latter stresses rather the individual as a member of a group, the members of which are all motivated towards the same end that is finding the solution to the problem

8 *Personality Judgments* The candidates were then instructed as follows It is an important part of the duties of a factory manager to be able to sum up other people and himself objectively, and if necessary, ruthlessly You should now write a thumb nail sketch of the other candidates and yourself, with particular regard to their personalities with respect to the position of trainee factory executive All of your reports will be anonymous and will not sway our judgment either against or in favour of any particular candidate ' They were seated at separate tables for this task in order to reduce any inhibitions that may have arisen from the close proximity of the persons being rated

The judgments made varied considerably in quality, and revealed varying willingness to unmask personalities The insight possessed by the candidates also appeared to vary considerably

9 *Report on the Proceedings* In his closing address the Managing Director requested the candidates to forward to him by mail a report recounting the proceedings of the day, and giving their impressions of what had occurred These reports provided an indication of each candidate s judgment, ability to write a report on factual occurrences, powers of observation, memory for details, and maturity in evaluating a situation

Evaluation of the Candidates

At the conclusion of each day s' observations the Managing Director and the Psychologist discussed and tentatively evaluated the candidates in terms of their suitability for the position in question Following the practice used in the evaluation of O S S candidates (5), they were not judged on their comparative levels on a number of traits, but they were discussed rather in terms of their weak and strong points as shown in the various situational tests conducted during the day

When the Rorschach tests had been scored and the reports received from the candidates a final selection conference was held All the available information and reports on the candidates were considered, with particular weight given to the group observation data, since the other data had already been used for screening purposes

Evaluation of the Procedure

A consideration of the validity of the group observation procedure involves two major questions (a) How well does it predict the ultimate success of the candidates? and (b) Does it add anything to the predictive power of the usual test battery plus interview?

It would be difficult to answer either of these questions in the absence of criteria provided by long range longitudinal studies However, in respect to question (b) it may be of value to compare the rankings made by the Psychologist after the vocational guidance interview with the overall rankings made at the completion of the selection procedure These are set out in Table 7 1

Candidates A and B—the selected candidates—would have been chosen as the first two choices without the group observation interview However, there are significant changes in the position of the other candidates, and it is possible that such changes could have occurred in the case of candidates A and B

As far as the individual items of the group observation sessions are concerned it is difficult to evaluate their separate contributions to the final result, as the day's proceedings have been viewed as a unit which develops progressively

Criticisms

1 The group situation used in selection is so variant from the actual situation as to be worthless as a basis for drawing inferences, if not actually misleading

TABLE 7 1

Showing Comparative Rankings of Candidates by the Psychologist After the Vocational Guidance Interview and the Overall Ranking After the Completion of the Selection Procedure

Candidate	*Rank Order after Voc Guid Interv*	*Overall Rank Order*	*Change after Group Observ*
A	2	1	+1
B	1	2	−1
C	5	3	+2
D	3	4	−1
E	10	5	+5
F	7 5	6	+1 5
G	9	7	+2
H	4	8	−4
I	11	9	−2
J	7 5	10	−2 5
K	12	11	+1
L	13	12	+1
M	6	13	−7

It is pointed out however that there is sufficient correspondence between the artificial' and the actual' situations to expect similar samples of behaviour For example, it would be expected that a candidate whose logic deteriorated as a result of emotional involvement in the leaderless discussion would show similar reactions in the everyday relationships with other factory executives

2 Inferences are not permissible from the ability of a candidate to lead the other candidates to his ability to lead a group of factory workers

This criticism is unavoidable in any form of selection excepting that of trial and error, and it is believed that the differences between the two social groups were constantly borne in mind by the observers

3 The group observation technique assumes consistency of behaviour from one situation to another (i e test retest reliability) without regard to temporary moods or reactions to unusual circumstances

However, if the candidate shows up badly during the observation, it seems a reasonable assumption that there will be occasions on the job when he will do likewise

Viewpoints on the Procedure

The Managing Director felt that the group observation procedure had given him an opportunity to participate fully in the selection procedure, and to obtain a preview of his potential employees' behaviour It had also eliminated much of the esoteric aura that has surrounded the work of the psychologist as seen by the layman

The reports submitted by the candidates showed that they too considered the procedure a particularly just one, eight of the fourteen stating this explicitly Several of them also revealed in their remarks signs of the self clarification which has been noted by other writers on this subject (This self clarification" can be compared to the insight which develops as a result of participation in role playing) One typical remark was "I shall always remember today as a day of enlightenment and experience in my life '

Summary

The problem of developing future executives for industry frequently requires a planned program involving the selection of potential executives from amongst com

paratively young and untried persons The usual methods of psychologically testing and interviewing candidates are limited by the difficulty of inferring social behaviour traits (such as dominance, cooperativeness, ability to persuade, stability in the face of emotional stress, sound judgment, etc)

During World War II the Group Situation Observation Method was devised, mainly by the British Army psychologists, to meet this difficulty in the selection of officers, and the metnod is now being applied to the selection of industrial and administrative executives An application of this technique has been described where the problem was to select two trainee factory executives for a small shoe factory The candidates were first screened by means of aptitude tests and an interview and were then divided into groups of four or five for observation The full day s procedure included a personal introduction by each candidate, a group Rorschach Test, an unstructured and a structured discussion period, and personality ratings by each candidate of the others

REFERENCES

1 Bridges H , and Isdell Carpenter, R , Selection of Management Trainees, *Industrial Welfare Personnel Management* 1947, Vol 19, 177–180, 315
2 Frazer, J M New type Selection Boards in Industry, *Occupational Psychology* 1947, Vol 11, 170–178
3 Garforth G I De La P , War Officer Selection Boards, *Occupational Psychology* 1945, Vol 14, 97–108
4 Gibb, C A , The Principles and Traits of Leadership, *Journal of Abnormal Social Psychology* 1947, Vol 42, 267–284
5 MacKinnon, Donald W , Some Problems of Assessment *Transactions of the New York Academy of Science* 1947, Vol 9, 171–185 (original not seen, quoted in *Psychological Abstracts* 1947, Vol 21 496)
6 Murray, H A Assessment of the Whole Person In Kelly, G A , *New Methods in Applied Psychology* Md , College Park, 1947 (original not seen, quoted in *Psychological Abstracts* 1947, Vol 21 496) See also, OSS Assessment Staff, *Assessment of Men* New York Rinehart & Co , Inc , 1948

Chapter II

THE APPLICATION BLANK

The application blank is a widely used tool in applicant screening, it follows the interview in frequency of usage Much of its use, however, is not harnessed to greatest efficiency This form can be constructed to obtain a considerable amount of personal history data related to job success Further, it is not necessary to make assumptions about such items All that is necessary is to establish a measure of successful job performance, then, analyze the personal characteristics of the employees to determine whether certain of these qualities are more often found in the successful workers and, conversely, to learn which qualities characterize the unsuccessful workers

A problem of procedure arises in deciding whether the personal history items should be established on the basis of the present characteristics of employees or on the basis of characteristics displayed when the employees were applicants To a large extent, which of these is used may be determined by the personnel policy of the company, and sometimes, unfortunately, on the basis of the kinds of records available

Such studies have been reported in books and periodicals in the field, and those of Kerr and Martin, Ohmann, and Guilford and Comrey illustrate clearly the

typical problems involved in validating application blanks It is worth remembering that any series of personal items printed on a form can be called an application blank, but this does not mean that it is necessarily a valid selection device

Kerr and Martin's study is based upon 244 employees with varied jobs in a large company The criterion of success was a rating by supervisors, the report indicates the correlations obtained The results indicate that certain items are useless and should be omitted from consideration in hiring Others are a little more meaningful and when combined can present a constellation to help guide in the hiring process

Ohmann's report yields insight into the complexities of establishing the criterion This problem in itself is a crucial one, for without establishing a measure of success it is impossible to predict who will be successful Ohmann determined thirteen valid items and, by establishing a numerical weighting system, obtained a rather high correlation between success on the job and the items on the application blank A worthy point to note in connection with this study was that he took one additional important step He conducted a follow up study This is always advisable, and helps insure against the possibility that the conclusions apply only to the subjects studied but not to other similar groups In this instance the results were still positive, and so it seems probable that the conclusions were not a result of any peculiarities of the first sample

Guilford and Comrey attempted to determine the applicability of this technique to a complex type of job Further, they were interested in extending the information usually sought in application blanks to material sometimes obtained in interest measurement tests They were not only concerned with the criteria to be established, but also indicated the types of difficulties one is likely to encounter in obtaining a representative sample of subjects

The study was obviously conducted with extreme care, and yet the results as reported by the authors are "disappointing" The main reason for including the Guilford and Comrey study is to re-emphasize that scientists do not always obtain positive results In fact, a preponderance of research leads to "disappointing conclusions" Many lengthy studies yield small and minute findings Each in a small way adds a grain or two to the total body of knowledge, however, and often this is all that can be expected

The layman with his incomplete understanding of research expects miracles and he sometimes innocently encourages the "quack" who is too willing to exaggerate If an industrial firm had paid for the Guilford and Comrey study, what would the reaction of the executive have been? If he said, "I paid and got nothing," he would be very wrong If he concluded that we now know that we can't do too much in this direction (for the particular job studied), then he would be in a position to attempt progress in another direction

The validation of the application blank often yields positive findings and so is a step in the right direction enabling employers to hire a greater proportion of applicants who will be successful employees

Prediction of Job Success From the Application Blank *

WILLARD A KERR
and
H L MARTIN

While considerable factual information is contained on the typical industrial personnel application blank, little information is now available to indicate the actual value of this information for predicting the probable job success of the applicant This study attempts to make a small contribution to existing knowledge on this topic by obtaining correlations between success on the job and such information items as sex, marital status, possession of telephone, street address (i e part of city), age birth place (in or out of state), children, dependents, height, weight, previous employment with company, insurance, recent illness or operations, number of personal references listed, organizations, hobbies, company acquaintances, education, and previous positions for 244 employees in the personnel, engineering, purchasing, production control, phonograph record manufacturing, electronic tube manufacturing, and warehouse departments of the Indianapolis plant of the RCA Victor Division of Radio Corporation of America

Success on the job measures for these 244 employees were obtained from supervisors and the raw merit ratings (split half reliability of the merit rating form was found to be above 75) from each supervisor were transformed into standard dichotomous scores which were then plotted with the information items to obtain tetrachoric coefficients of correlation between job success and these variables These are presented in Table 8 1 Correlations significant at the five per cent level or better are set in boldface type

On the basis of the highest correlations, eleven items were scored, check list fashion, and the total scores were correlated with job success to obtain a coefficient of 35 ± 04

Although all these findings should be accepted as tentative, it is possible that some of the findings may be found to apply to most departments of work in general industry Analysis of item predictive value for various types of work was not attempted here because of the limited number of cases

Marital status has the greatest relationship with the criterion for these cases Area B street address correlates positively with job success while Area A address correlates negatively, this is regarded as surprising since the large area of above-average socioeconomic status in the city is in Area A Number of children and number of dependents, while regarded as important by most personnel workers for obvious social reasons, do not correlate significantly with job success Height and weight seem to be of little general predictive value, although obesity tends to be a definite handicap for men Former employment in the company seems to be an asset, but possession of insurance and recent illness or operation appear relatively unrelated Listing of an excessive number of personal references, hobbies, or previous positions is negatively related with the criterion, but membership in organizations and special education are positively related

It should be emphasized that these correlations are low and the findings possibly may apply only to the workers measured Nevertheless, it is interesting to note that approximately ten per cent of the variance in job success of these 244 employees is accounted for by "autobiographical factors reported in the original applications for employment Such check list autobiographical scores may make a highly useful

* Reprinted from *Journal of Applied Psychology,* Vol 33, No 5, October 1949

TABLE 8 1

Correlations between Job Success Ratings and Personal History Items

Item		r
Female sex		− 16
Marital status	single	− 18
	married	30
	divorced	− 05
Telephone number (possession of)		07
Street address	Area A	− 22
	Area B	23
	Area C	15
	Area D	− 11
Age		08
Birthplace (in same state as plant)		15
Number of children		00
Number of dependents		00
Height of males		− 12
Height of females		05
Weight of males		− 27
Former employee of same company		22
Holds insurance policy		06
Recent illness or operation		00
Number of personal references listed		− 17
Number of organizations in which membership is held		23
Number of hobbies		− 18
Number of company acquaintances		− 09
Education	special training	15
	college	− 01
Number of previous positions		− 22

addition to the total predictive test battery Naturally they should not be weighted more heavily in determining selection than their relative contribution to determination of job success variance indicates In order to maintain the validity of the autobio graphical scoring key, it should be revised periodically according to results of routine revalidations

Better results may be obtained in select ing for a specific job with this device than when using it to hire for the entire plant Manson (1), for example, found a coef ficient of correlation of 40 between the weighted scores on an application blank and the production records of life insur ance salesmen, and Ohmann (2) obtained a correlation of 67 between his blank and the earnings of paint salesmen

Summary

1. Most of the items on a typical in- dustrial personnel application blank are easily quantified in check list fashion on the basis of a previous item validation study against a job success criterion

2 In this study, when the original ap plications of 244 employees were scored check list (unweighted) fashion with a validated key, the check list raw scores were found to correlate 35 with the super visory merit ratings of job success

3 Since in this study the application blank accounts for approximately ten per cent of the variance in job success of an extremely heterogeneous (almost "run of the employment office') group of em ployees, it seems reasonable that the ap plication blank or a systematic auto biographical inventory should become a standard part of the psychometric battery in industry

4 In view of the facts that background factors change in predictive significance with time and their significance also is altered by changes in the business cycle, the industrial psychologist should revali date such an instrument periodically

5 When validation keys are developed

for specific kinds of employees or job families, more substantial correlations are likely to be obtained both with the job success criterion and the tenure criterion

REFERENCES

1 Manson, G E, What the Application Blank Can Tell, *Journal of Personnel Research* 1925, Vol 4 1–28
2 Ohmann, O A, A Report of Research on the Selection of Salesmen at the Tremco Manufacturing Company *Journal of Applied Psychology* 1941, Vol 25, 18–29

A Report of Research on the Selection of Salesmen at the Tremco Manufacturing Company *

O A OHMANN

Industrial psychology has made most headway in the large, well established business organizations where a long time research point of view toward personnel problems has been developed, and where there are enough employees in any given occupation to make quantitative measurement both possible and profitable In Cleveland there are 50 firms employing 500 or more workers, 108 firms with 200–500 workers, and 157 firms with 100–200 employees This report may suggest some possibilities for psychological research in a relatively small organization, provided a continuing personnel research program has been established A long time approach to the problems of selection is imperative here because of the small samples available for study at any one time Research attacks on various fronts may be made simultaneously, and measurements and records accumulated over a period of years for later validation This report will be in the nature of an outline of the program of research on the selection of senior salesmen which has been initiated at Tremco Manufacturing Company Progress made in various areas of study will be indicated and tentative findings summarized

The Tremco Manufacturing Company operates plants in Cleveland and Toronto for the manufacture of materials used in the construction and maintenance of buildings—such as materials for caulking, pointing, waterproofing, roofing, painting, flooring They employ 75 salesmen located in territories in the eastern half of the United States and all of Canada These men call on plant superintendents, schools, apartment owners, hospitals, large property owners, and on contractors and architects They do not as a rule sell to individual home owners, nor to dealers or jobbers Each sale is creative in the sense that it requires the diagnosis of a particular problem on a building, and the specification of materials and methods for making the repair The importance of an effective sales force in such a business and the high cost of turnover indicated that the problem here was one of improving methods of selection and training The research approaches which have been made to the problem of selection may be outlined as follows

1 *Study of criteria* The problem here was to determine the best measure of performance of salesmen, so that predictive items and instruments could later on be validated against this criterion Fortunately territories had been established on the basis of equal sales potentials, as determined by various business indices The experimental group for the study of criteria consisted of 30 salesmen who had been with the company for more than two years, and whose records were sufficiently complete for the year 1937 to permit the intercorrelation

* Reprinted from *Journal of Applied Psychology*, Vol 25, No 1, February 1941

of the following measures of job performance

Sales volume for 1937
Average number of calls per day
Number of years worked at Tremco
Salesman s net commission earnings for 1937
Average number of sales per month
Average size of order
Average number of new accounts per month
Average sales volume per year for length of time employed
Sales volume for the first six months on the job
Trend of sales volume over a period of years
Amount of allowances to customers
Amount of returned merchandise
Classes of trade called on
Classes of products sold

This study resulted in establishing the "salesman s net commission earnings as the best single criterion The investigation also yielded other interesting by products For example, number of calls per day" correlated negatively with all other measures of success, including the number of sales Salesmen on the job a long while tend to open very few new accounts, although there is a possibility of doing this Consistency of performance of salesmen from year to year was shown by a correlation of 92 between 1937 sales and ' average sales volume for length of time employed ' However, sales volume during the first six months' did not correlate highly (44) with later annual sales volume This last finding suggested the advisability of a more thorough investigation of the relation between sales during the first months of employment and later success

In order to determine whether sales during the first few months might be used as a basis for early elimination of men who are likely to fail later on, the records of a group of 65 salesmen were studied Cumulative sales volume for the first, second, third, and fourth months were correlated with year end sales volume Since these correlations range between 20 and 30 it was concluded that too much stress had been placed on the sales volume of new men Subsidies to new men (in the form of overdrafts permitted) were entirely uncorrelated with year end sales volume Further analysis showed that sales volume during the first four months could be used to predict the most serious ultimate failures but that average or high sales during the first few months did not indicate a corresponding degree of later success Company policies with reference to subsidizing new salesmen have been adju ted in line with these findings

2 *Study of the application blank and personal history data* A four page application blank with 31 items had been in use for eight years Each of these items was evaluated for an experimental group of 48 salesmen whose earnings records were available for the full year of 1938, and who had filled out the application blank at the time of their employment Six of the 31 items had to be thrown out because of the impossibility of objective scoring The remaining 25 items were evaluated on the basis of differentiating between the upper and lower half of the sales force when 1938 earnings were used as a criterion Twelve of these 25 items were found to be totally invalid on this basis The 13 valid items remaining were then given scoring weights on the basis of the percentage of men giving each possible answer who fell within the limits of the upper half of the sales force in terms of 1938 earnings For example, for the item of 'Age" the scoring weights listed below mean that 70 per cent of the men who were under 40 years of age when employed were later in the upper half of the sales force in earnings, while only 20 per cent of those between the ages 40–44 were later above average in success The scoring weights listed above are regarded as tentative, and as applying only to salesmen at Tremco Manufacturing Company

Total scores on the 13 items obtained for each of the 48 salesmen in this experimental group correlated 67 with the 1938 earnings criterion Critical total scores were then set to indicate a "danger zone" in employing new salesmen In actual practice the sales managers did not adhere rigidly to these critical scores This was in line

Scoring Weights for Personal History Items

Item	Score
1 Age	
50	4
45–90	5
40–44	2
up–39	7
2 Height	
72′–up	7
70 –71 9	5
69″–69 9	4
up –68 9	3
3 Marital status	
Married	5
All others	3
4 No of dependents	
4 or more	0
3	3
2	6
1	7
None	3
5 Thousands of Ins	
10 or more	5
5 to 10	6
1 to 5	3
None	6
6 Amount of debts	
None	4
Current	6
$500 or more	5
7 Years of education	
Grades 1–8	6
9, 10, 11	3
12, Col 1	6
Col 2 3	0
Col 4, more	5
8 Number of clubs	
None	6
One	4
Two	6
Three, more	3
9 Years on last job	
Less than 1	5
1 to 1 yr 11 mo	1
2 to 2 yrs 11 mo	3
2 to 3 yrs 11 mo	6
4–5 yrs 11 mo	8
6–9 yrs 11 mo	10
10 or more	5
10 Experience in maintenance	
None	3
Any amount	6
11 Average No Years on all previous jobs	
1 – 2½	3
3 – 6	5
6½–10	8
12 Average monthly earnings on last regular job	
Up to 150	5
150–199	4
200–249	8
250–349	1
350–399	5
400–up	6
13 Reason for leaving last regular job	
Still employed	10
Job discontinued (depression) (Co folded) (also illness and circumstances beyond man's control)	7
To better self (positive reasons)	5
Was let go—dismissed (but if because of conflict with management score as negative reason)	4
Negative reasons (friction)	2

Critical Score=62

The experience of the company is that 70% of those scoring above 62 are still working while only 30% of those scoring below 62 are still employed

with the established policy because of the recognized unreliabilities of scoring weights based on such a small sample

During 1939 and 1940, 65 senior sales men were employed A follow up study has just been made to determine what the effect of strict adherence to the critical scores would have been with this group Since many of these men had been recently employed, it was possible to establish only a rough two fold criterion—those who had actually failed and were no longer employed, and those who were still working The results are as follows

Of the 65, 33 were still working, 32 had failed —Ratio 1 to 1
If critical scores had been followed rigidly, only 25 men would have been employed, of whom 20 are still working, 5 have failed —Ratio 4 to 1
40 of the 65 would *not* have been employed
13 of these are still working, 27 failed —Ratio 1 to 2
Of these 13, however, 8 are regarded as probable failures by the end of the year

These 13 items, together with some new experimental questions, constitute the new Personal History Record form Critical scores on this form are now regarded as crystallizing the experience of the company and rendering it usable

Prediction of Proficiency of Administrative Personnel From Personal-History Data *

J P GUILFORD and ANDREW L COMREY

[The authors are] indebted to Dean Emery E Olson and Professor John M Pfiffner of the School of Public Administration, who made this study possible by allotting funds from a research grant also to Dr Pfiffner for his continued interest and support, and to Dr Paul E Webb, E C Wills, Raymond E Pollich, and Elizabeth Sands of the Los Angeles City Schools, who cooperated in the study

Attempts to select administrative personnel by means of tests have not been as numerous or as successful as those in connection with other types of personnel This can perhaps be attributed to the fact that administrative ability represents a varying complex of many different traits, few of which are easily isolated The investigation reported here represents an attempt to apply a biographical data technique to the task of measuring administrative ability It was believed that the biographical data method had not previously been given an adequate trial in this particular field of personnel selection

The encouragement to try this method stemmed largely from the success with which it was employed by the United States Army Air Forces (3, chapter 27) in the selection of pilots and other air crew members Although the resemblance between these tasks and those of administrative personnel is not great, both types of tasks are markedly complex If the method could be applied to one complex task as a whole, it seemed reasonable that it might be successful for another

Of particular interest in connection with the Army Air Forces studies was their investigation of the effect of directions upon the validity of the biographical data test It was found that standard directions with no mention of penalty for falsification gave results which were equally as valid as directions threatening severe punishment for falsifying the response Directions which encouraged laxity in answering the questions gave distinctly lower validity These results are important in any consideration of a biographical data test for selecting employees, since the opportunity for falsification is great in this type of test

Studies with life insurance company employees have shown that biographical data items are useful in predicting success in certain types of jobs Several attempts have been made to utilize other printed tests to measure administrative ability with at least some degree of success Thurstone (8)

* Reprinted from *Educational and Psychological Measurement* Vol 8, No 3, Autumn 1948

found that the Social Scale of the *Allport Vernon Scale of Values* differentiated among certain federal government administrators Mandell and Adkins (5) obtained validity coefficients greater than 60 for a top management group of executives on such tests as the linguistic section of the *American Council on Education Psychological Examination* a *Civil Service Commission Current Events Test,* selected items from an *Interpretation of Data Test* of the Progressive Education Association, an administrative judgment test, and an agency organization and-personnel test Mandell (6) reported that the United States Civil Service had obtained promising results with the *Kuder Preference Record* He believes that there is some evidence in favor of tests of mental ability and for interest inventories Strong (7) presented some evidence to suggest that differences in interests between administrators and others may be expected Achard and Clarke (1) reported definite indications that the measurement of interests can contribute to the selection of supervisory personnel They used the *Vocational Interest Blank for Men (Revised), Form M* Scale CFS, with success and also the *Otis Self Administering Test of Mental Ability Higher Examination Form A* Uhrbrock and Richardson (9) found that mental ability items were useful in differentiating between good and poor supervisors A few biographical data items were also helpful These items related to age, schooling, confidence in blue print reading ability, and military service Mitchell (4) used biographical data in selecting sales managers He found that better educated sales managers were more successful, but other items of biographical information did not prove to be effective

Development of a Biographical Data Inventory

As it has been used in the past, the familiar application blank represents an attempt to use biographical data, usually informally and unsystematically, in selection Most of the research which has been done in various fields of personnel selection with such data has utilized an open end type of question and a somewhat limited area of questioning In the present investigation, the intent was to utilize a large number of questions covering a wide range of biographical information In addition it was decided that the questions should be put in multiple choice form so that adequate statistical analysis of the results could be carried out This feature of the test also provided for the use of standardized answer sheets and machine scoring In this way much of the labor was eliminated, both for the subjects and the investigators

The biographical data booklet in its final form was a fourteen page printed booklet containing 150 multiple choice items The items may be loosely classified into four different types (1) childhood background and family life, (2) professional preparation, (3) health, (4) interests and (5) early signs of leadership

The items of the first group involved information concerning occupations, interests, and social activities of the parents, social habits as a child, and happiness of the early home environment It was believed that the character of the home environment might do much to influence the development of those traits which are important in administrative ability A sample item of this type was

18 During most of the time before you were 16, you lived

18—A With both parents
18—B With one parent
18—C With a relative
18—D With foster parents or non relatives
18—E In a home or institution

The second group included items concerning the amount of education completed, types of subjects studied, scholarship and awards, major academic interests, and previous types of employment Previous work with biographical data had indicated that amount of education was of value in the selection of administrative personnel, so it was hoped that these as well as similar items might be significant A sample item of this type was

47 As a college student, you were

47—A A Phi Beta Kappa
47—B A Sigma Xi

47—C Graduated with honors (But not 47—A or B)
47—D A good student (But not 47—A, B, or C)
47—E An average student

The third group of items concerned the health of the individual, both as a child and as an adult The administrator is gen erally believed to be a person of excellent health and considerable energy It seemed likely that individuals indicating a history of less than average health might prove to be poor administrators A sample item of this type was

25 Between the ages of 12 and 21, how often were you sufficiently ill to re quire hospitalization?

25—A 0
25—B 1
25—C 2
25—D 3
25—E 4 or more times

The fourth group of questions included a large number of items relating to inter ests, types of recreation enjoyed, reading habits, motion picture and entertainment preferences, and social habits In view of the suggestion of previous research that interests may be of value in the selection of administrators, it seemed important to include items of this type A sample item was

141 What type of radio program do you prefer?

141—A Classical music
141—B Dance bands
141—C News commentators
141—D Plays
141—E Comedians

The fifth, and final, group of items was devoted to questions concerning early signs of leadership Since administrative work involves leadership in most cases, it was believed that individuals who were leaders in their childhood and adolescence might later become good administrators A sample item was

35 How often were you a leader of your childhood "gang" activities up to the age of 12 years?

35—A Always
35—B Frequently
35—C Occasionally
35—D Seldom or never
35—E (You were not a member of a group, or you can t remember)

Many other heterogeneous types of questions were included which cannot properly be placed in one of the four men tioned groups For example, one item was

101 Your own personality most re sembles that of your

101—A Father
101—B Mother
101—C Brother
101—D Sister
101—E (None of these or you don't know)

Certain types of items were omitted be- cause it was felt that they might prejudice the subjects against the research and hence do more harm than good Questions per- taining to race, religion, and very personal matters were not included

Item Validation Procedures

The original sample included all of the regular principals and vice principals of the Los Angeles City Schools These fell into three groups, elementary, junior high, and senior high school principals In prep aration for the administration of the test, one of the investigators talked before most of the principals in the elementary and junior high school groups to explain the purpose of the project, and to assure them that the information obtained would not be utilized by the school system for pur poses of evaluating individual principals It did not prove possible to offer this same type of orientation for the senior high school principals

The biographical data booklets were en- closed individually in an envelope together with an instruction sheet, a standard answer sheet, and an electrographic pencil One of these envelopes was then distributed via the official inter school mailing system to each one of the principals to be filled out and returned The principals were also in- structed to return the booklets in sealed

envelopes through the inter school mail It would have been most desirable to have had the principals fill out their answer sheets in supervised groups This kind of administration of the test was regarded as not feasible by the school authorities

Returns from the senior high principals numbered only 37, which constituted only a fraction of the total number in that group For this reason, the results for that group were not analyzed The elementary and junior high groups were combined to be treated as one group Of this combined group, 328 principals returned booklets, which constituted a return of 93 per cent A few of those who did not return booklets were ill, the remainder either refused to turn them in, or ignored requests to do so It was felt, however, that the remaining seven per cent of the cases not included in the analysis would not materially alter the results

The 328 cases which constituted the final sample were subdivided into Groups A, B, and C Group A was composed of 122 female principals selected at random, Group B included the remaining 123 females, Group C was composed of the 83 males included in the total elementary and junior high sample The female cases were separated into two groups so that a comparison could be made between the results of the analysis in one group and those in the other to obtain evidence as to whether any significant correlations obtained were not due to sampling errors It was also planned to make a cross validation study, i e, derive a scoring key for valid items in one group and estimate the validity of the total score in the other group In view of the item analysis results, this step proved to be unnecessary The men were put in one group because there were not enough of them to make up two groups

The answer sheets filled out by the principals were divided into six groups, according to the division of the principals into high and low halves for each of the Groups, A, B, and C Then, a scoring machine tally was made of every response, for each alternative answer to every question for each of these six groups For example, in question one, with five alternate answers, we might find that 30 individuals in the high group and 10 individuals in the low group of Group A had marked the space corresponding to alternative 1 A For Groups B and C, different tallies would occur

With 150 questions, each with five alternative answers, a total of 750 categories were tallied for each of the six groups Then, a correlation could be computed for each response for Groups A, B, and C separately Thus, 750 correlations were possible for each of the three groups, or a total of 2250 correlations for the entire study On the basis of visual inspection of the differential frequencies, plus previous calculation of the difference in frequency required for approximate significance, many of the correlations could be eliminated as being definitely below the level of significance For the remaining ones, phi coefficients were computed, using an abac provided by Guilford (2)

The Criterion

One of the greatest difficulties in studies of the selection of administrative personnel is to decide who is and who is not a good administrator This was perhaps the greatest difficulty and the weakest point in this research The problem with which we were confronted was to divide each of the three Groups A, B, and C into two halves, higher and lower, with respect to administrative ability

The only information available for this purpose consisted of the periodic ratings made by the school system for purposes of promoting principals to the next higher salary groups This particular problem was studied for some months prior to the mailing of the booklets to see whether the ratings were of sufficient reliability to warrant a study It was decided that the ratings exhibited sufficient *reliability* to enable us to carry out the research, but no determination could be made with respect to the *validity* of these ratings

The ratings of each principal are generally made independently by three or more superintendents There is a definite possibility that the basis upon which superintendents rate individual principals may be colored by something other than their administrative ability

Many difficulties arose in connection with the evaluation of these ratings The elementary principals were rated in groups generally by means of a rank order technique, in which each of approximately four superintendents arranged in rank order a stack of filing cards each of which held the name of a principal being considered for promotion The secondary or junior high principals were generally rated on a five category graphic rating scale with a spread of 100 points for each category These categories included such traits as success in community relations, ability as a manager, success with personnel, and the like A study of the intercorrelations between traits showed that a pronounced halo effect was operating, hence an average of the five different categories might just as well be used

The difficulties with these ratings included the following ones, among others which we may have overlooked The principals in Group A for example, had not all been rated by the same superintendents, hence no satisfactory basis existed for comparing each principal with every other principal Furthermore, some principals had been rated several times over a period of years while others had been rated only once, or not at all Further, a comparison was made necessary between principals who had been rated on a graphic scale and those who had been rated by a rank order method

The problem of determining the reliability of these ratings proved even more of a dilemma In order to estimate the reliability of the ratings, either different raters on the same individuals should be compared, or the same rater at different times, or both These procedures require having data for all principals rated by the same superintendents on different occasions Since only a few of the principals in a group had been rated in such a manner, it was very difficult to make a determination of the reliability

In an attempt to get around these difficulties, the following simple procedure was used A list of all the principals in each group was made For every occasion on which he had been rated, the principal was given a relative position in comparison with the other principals rated along with him That is, on the basis of an average of all his ratings, whether rank order or graphic rating scale, every principal in the particular group being rated at the moment was given a rank order number This number was translated into a per cent position by the following formula

$$\text{Per cent Position} = \frac{100(R - 5)}{N},$$

in which R is the principal s rank order number and N is the number of individuals with whom he is being compared in that particular group Each per cent position was then converted into a score from 0 0 to 10 0 in which 5 0 is approximately the mean It was assumed that the individuals in each group were normally distributed with respect to the qualities rated and that the mean proficiencies of groups were actually much alike

For those principals who had not been rated previously, special ratings were obtained and treated in the same manner described above After all this manipulation of data, a list was available in which every principal had a numerical value beside his name which was roughly comparable with the numerical values of every other principal In order to obtain one final representative score, an average was computed of all the scores for a principal, so that finally, one score for each principal was obtained

In order to make an estimate of reliability, it was arbitrarily decided to include all principals who had been rated at least twice Then, the first and last tabulated rating for each of these principals was taken A phi coefficient (corrected for continuity) between these first and last ratings for 242 cases was 66, which was significant beyond the one per cent level

In each of the three groups, A, B, and C, a median of the composite rating scores was determined On the basis of these medians, the principals in each group were divided into higher and lower halves

Results and Conclusions

Frequency distributions of the phi coefficients of significant size are given below

The levels of significance given were determined by reference to chi square values, converted to corresponding phi coefficients Although the uncorrected phi coefficients were used in determining their significance, the coefficients in this table, and below it, have been corrected for continuity in one variable[1]

The 86 statistically significant responses listed in Table 10 1 were distributed among the three groups as follows (1) 24 responses were significant in Group A only, (2) twelve responses were significant in Group B only, (3) thirty responses were significant in Group C only, (4) four responses were significant in the same direction in Groups B and C, (5) one response was significant in the same direction in Groups A and C, (6) two responses were significant in the opposite direction in Groups B and C, (7) one item was significant in the opposite direction in Groups B and C, (8) no response was significant in the same direction for all three groups, (9) no response was significant in the same direction for Groups A and B, the two female groups, and (10) a total of 57 items were involved in the 86 significant correlations for the three groups combined

An analysis of the 57 significant items by groups reveals (1) fourteen were significant in Group A only, (2) nine were significant in Group B only, (3) twenty were significant in Group C only, (4) five items were significant in Groups A and B Of these five items, the alternative response results agreed substantially in one, disagreed in two, and neither agreed nor disagreed in the other two, (5) five items were significant in both A and C groups Of these, three were in agreement in the two groups, and two responses neither agreed nor disagreed (6) three items were significant in Groups B and C The responses in these items were significant in the same general direction in both groups and (7) one item resulted in the samc trend in all three groups

Responses within items which reached the one per cent level of significance in Group A were (1) participation in mild sports (golf, hiking, etc) occasionally, phi = − 33, (2) participation in mild sports (golf, hiking, etc) seldom, phi = 35, (3) participation in collecting (stamps coins, antiques, etc) frequently, phi = − 35, (4) between the ages of 12 and 18, belonging to an organized group of children of own age without adult sponsorship, phi = 31, (5) other than school work, reading during a large part of your free time

TABLE 10 1

Distribution of Significant and Marginal Phi Coefficients Between Responses to Items and the Criterion of Administrative Proficiency

Phi	*Group A*	*Group B*	*Group C*	*Total*
60–64			1	1
55–59				0
50–54			1	1
45–49			1	1
40–44			2	2
35–39	2	1	4	7
30–34	6	2	19	27
25–29	16	8	7	31
20–24	8	8		16

Group A N = 122 05 level = 23 01 level = 29
Group B N = 123 05 level = 23 01 level = 29
Group C N = 83 05 level = 28 01 level = 35

[1] This type of correction rests on the assumption that the criterion is actually continuous and that each response represents operationally a point distribution The corrected coefficient should be numerically equivalent to a point biserial r

between ages of 12 and 18, phi = − 31, (6) participation in making speeches frequently, phi = 31, (7) associating socially as a general rule with people younger than yourself, phi = − 29

For Group B, the following responses were significant at the one per cent level (1) if the cost were the same, you would prefer to travel across country for pleasure by private automobile (rather than by bus, airplane, etc), phi = − 38, (2) as a child, you confided most in a brother or sister, phi = 31, (3) you succeeded well in history as a college or school subject, phi = − 30

Responses reaching the one per cent level of significance in the male group were (1) prior to the age of 21, you lived most of your life in a large city (over 500,000), phi = 46, (2) you learned to swim at age 10, or below, phi = 35, (3) between the ages of 12 and 18, you belonged to the boy scouts, phi = 40, (4) you succeeded exceptionally well in history as a school or college subject, phi = 63, (5) you succeeded exceptionally well in psychology as a school or college subject, phi = 35, (6) participation in gardening occasionally, phi = 50, (7) participation in travel frequently, phi = − 44, (8) you are occasionally interested in making things, shop work (without necessarily having done so), phi = 38, (9) you are frequently interested in camp counselling, YMCA or YWCA work (without necessarily having participated), phi = − 38, (10) your present weight is 170 to 189, phi = 37

From the eight items which gave significant results in the same general direction for at least two of the three groups, the following information is suggested (1) the successful administrator's father generally employed more than five people, (2) confiding during childhood in the mother is not so auspicious as confiding in a brother or sister, (3) a child who when ill is put to bed, but with medication is more apt to be a successful administrator than if he had a physician called, had only home remedies, was merely kept at home, or had no special attention, (4) belonging to an organized group of children between the ages of 12 and 18 is worth while, (5) succeeding well in history as a school or college subject seems bad, but succeeding exceptionally well seems very good, (6) succeeding either well or exceptionally well in psychology as a school or college subject is not desirable for a female administrator, whereas succeeding exceptionally well in psychology is good for a male administrator, but only succeeding well is undesirable, (7) individuals who would like to have four or more children are not likely to be good administrators, and (8) those who were 40 to 50 years of age tended to be rated better as administrators, while those over 60 tended to receive poorer ratings

In evaluating the significance of the results on an over all basis, two approaches could be adopted (1) a comparison of the number of significant *responses* in each group and in all three groups combined with the number of responses which would be expected to reach a standard of significance by random sampling among responses that actually correlate zero with the criterion, (2) a comparison of the number of *items* which contained at least one significant response with the number of items expected by chance, considering each group separately Neither of these approaches is completely satisfactory, as will be pointed out

The first approach reveals a total of 86 responses out of 2250 to correlate significantly at or beyond the five per cent level Slightly more than 100 significant responses would be expected if the normal sampling situation could be applied in this manner Since the respondent is forced to choose only one of five alternative responses to an item, the 2250 judgments were by no means independent For this reason, the first approach does not appear to be applicable Because of the interdependence of responses, mostly in the form of *negative* correlation, a smaller number than 2250 should be used as the base, how much smaller is not known

The second approach, evaluating the significance of the items as a whole, is probably the better of the two methods If any response in a particular item demonstrates significance, that item as a whole may be considered significant With a total of 150 independently answered items, ap

proximately eight would be expected to yield correlations which equal or exceed the 5 per cent level of significance Actually, 25 items were significant by this criterion in Group A, 18 in Group B, and 29 in Group C This would suggest that a test composed of the significant items would have a predictive value significantly better than chance While by this procedure there would seem to be a marked excess in the number of significant items for each group, there is at least one qualifying thought to detract from this conclusion The items themselves are probably not independent Just how the item intercorrelations may affect the number of items meeting the criterion of significance by chance effects alone is not clear Since the item intercorrelations are probably positive (in the sense of the association of good features), chance factors might be expected to increase the dispersion of phi coefficients and thus to produce more than the normally expected number of significant items

We are consequently forced to give attention to the comparison of item phi's obtained from different groups There was not sufficient agreement among the results from the three groups with respect to the items which showed significance to warrant much optimism Only eight of the 57 significant items showed agreement in at least two of the three groups

The results of this study do not bear out previous findings with respect to the value of present interests for the measurement of administrative ability There is a slight indication that early environmental data may be sufficiently useful to warrant further investigation The items used did appear to select male school administrators slightly more effectively than female administrators This might indicate that female administrators in schools are a more homogeneous group than the males, or that biographical influences are not as uniformly effective with respect to administrative qualities in women

A number of possible conditions might have accounted for our failure to obtain promising results First, it might be that the items which were utilized were not sufficiently inclusive to reveal the areas of biographical data which would give the best results Secondly, the school principals who participated might not have been truthful in giving their answers It was mentioned earlier that one group had to be eliminated from the study because so few principals returned their booklets If the other two groups merely returned the booklets without giving careful attention to their answers, the results would be invalidated In view of the findings in the Army Air Forces study previously mentioned with respect to truthfulness of responses, lack of veracity does not appear to be a likely hypothesis, however

A further consideration of importance is that of criterion validity If the ratings upon which the division of the principals into higher and lower groups was made did not happen to represent real variations in administrative ability, it would not be surprising if negative results occurred

The reliability of the criterion was sufficiently low to result in obscuring a definite relationship, if it happened to be one of low correlation In such an instance it would be impossible to conclude that no real relationship exists

Assuming that none of these considerations actually was responsible for the disappointing results obtained in this study, the problem of generalizing to other types of administrators and even to other groups of principals remains to be considered In order to apply these results to other groups, we must assume that such administrators utilize the same types of traits for success in their work as the school principals, as rated, in this sample This problem cannot be solved except through further research with the biographical data technique using other groups of administrators

Realizing that so many variables are involved in the results reported here makes it difficult to draw any conclusions A review of the results, however, definitely points to the conclusion that the biographical data method has questionable value for the selection of school administrators A reasonable presumption would be that this conclusion extends to similar information obtained in application forms or in interviews, or at least casts suspicion upon that information

SUMMARY

An attempt to develop a personal history inventory for the selection of school administrators is described A 150 question, multiple choice type, inventory was given to more than 300 school principals and vice principals An item analysis was conducted in which responses to various items were correlated with success as a principal, measured by promotional ratings Although a significant number of correlations was obtained separately in each of three groups of these principals, good agreement between the groups was obtained with only eight items Possible explanations of the results were reviewed It was concluded that the biographical data method has only limited promise of usefulness for the selection of school administrators

REFERENCES

1 Achard, F H and Clarke, Florence H, 'You Can Measure the Probability of Success as a Supervisor *Personnel* Vol 21, 1945 353–373
2 Guilford, J P The Phi Coefficient and Chi Square as Indices of Item Validity, *Psychometrika* Vol 6 1941, 11–19
3 Guilford, J P (Ed), *Printed Classification Tests* Army Air Forces Aviation Psychology Program Research Reports Report No 5 Washington D C Government Printing Office 1947
4 Mitchell J H, An Experiment in the Selection of Sales Managers *Occupational Psychology* Vol 12 1938, 308–318
5 Mandell M M and Adkins D C The Validity of Written Tests for the Selection of Administrative Personnel *Educational and Psychological Measurement* Vol 6, 1946, 293–312
6 Mandell, M Testing for Administrative and Supervisory Positions, *Educational and Psychological Measurement* Vol 5 1945, 217–228
7 Strong, E K Jr Interests of Public Administrators, *Public Personnel Review* Vol 6, 1945, 166–173
8 Thurstone, L L *A Factorial Study of Perception* Chicago University of Chicago Press, 1944
9 Uhrbrock R S and Richardson, M W Item Analysis The Basis for Constructing a Test for Forecasting Supervisory Ability, *Personnel Journal* Vol 12, 1933, 141–154

Chapter III

TRAINING

It is axiomatic that an experienced worker or foreman is not necessarily a competent teacher, for in addition to knowing a job, one must also know how to teach Industry has not generally recognized this principle and faulty training often has occurred as a result

At other times, industry has attempted to avoid the problem of training by hiring experienced operatives and assuming that somehow correct training was previously obtained For greatest efficiency, however, it is best to face the problem of training directly This requires that all workers be trained by people who know how to train

Prior to any training, a job analysis should be conducted so that the important components of the job are determined Further, the difference in performance between good and marginal workers should be established Effective training demands that emphasis be placed upon correct methods, although recognizing the existence of individual differences suggests that acceptable modifications may be made to standard performance A good teacher is aware of this as well as of many other established psychological principles that have resulted from the extensive laboratory study of learning

The primary reason for including this chapter is to make clear that effective training is more than arm chair philosophy and self evaluation The studies of Lindahl, Katzell, McGehee, and Kellogg do not cover the full range of problems confronting one faced with introducing and maintaining effective training programs Each pinpoints a specific problem, however, and is worthy of study for its careful methodology and conservative claims based upon available data

Lindahl's report indicates the importance of analysis of movement as a basis of a training program Since the job involved required precision hand-foot coordination, a mechanical device was constructed to record movement analysis This apparatus led to the determination of the correct foot pattern to be used Lindahl measured the beneficial effects of training new operators correctly He shows how it is possible to extend such a program to improve performance of operators already on the job

Training can benefit all workers, from operatives to supervisors Katzell presents data to show the benefits of a thirty two hour course given to supervisors who already had an average of 18 6 years of experience The criteria used were changes in attitude toward supervision and attitudes toward the value of the course Katzell is mindful of the limitations of these criteria, but at least has offered more than self-evaluation

The McGehee study is important because it attempts to determine how early in a training period one can predict the amount of training required to reach average production The data indicate that fast learners reach standard production sooner Before the reader jumps to the conclusion, "Well, that is obvious," it would be wise to consider that an arm-chair controversy persists over whether the slow learner is the surer learner or whether the fast learner eventually knows more

According to McGehee, it is possible, with varying degrees of statistical accuracy, to predict in early stages of the training period those who will be most efficient so that a decision concerning who shall be continued in training can be made This study is, therefore, closely concerned with training costs

Ever since Bryan and Harter reported a learning curve for telegraphy in 1899, many researchers have been interested in graphically presenting the learning process Kellogg's objective method to plot the learning curve for flying an airplane and the results he obtained indicate that such a curve is similar to the ones obtained in the development of other skills Objectivity in determining the learning process was introduced by use of the pilot response recorder, the graphometer, and the weather-control technique Such equipment and procedure, when coupled with a standard type flight, make it totally unnecessary to estimate subjectively the learning process in flying, it allows for awareness of several kinds of errors and accordingly can be an aid in the training process

Movement Analysis as an Industrial Training Method*

LAWRENCE G LINDAHL

Based upon a thesis submitted by Lawrence Gaylerd Lindahl to the Faculty of Purdue University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy, October 1944 Acknowledgment is due Dr Joseph Tiffin and Dr C H Lawshe who jointly directed the research and to James R Brock who made the study possible in industry

Many industrial jobs which involve co ordination of hand and foot movements in operating machines present difficult training problems Typical of such jobs is con tact disc cutting The company producing contact discs experienced considerable dif ficulty in training new operators because the majority did not complete the training and for those who did, learning was slow and uncertain Preliminary investigation with experimental apparatus indicated that part of the difficulty encountered in train ing new operative personnel was due to the failure to identify the form of cutting movement necessary for 'getting the feel' of satisfactory performance of the job

The cutoff machine slices thin discs (e g, 020" × 150' diameter with ± 002 on both dimensions) from various sizes of tungsten rods with a rubber bonded abra sive wheel 015 of an inch thick and six inches in diameter The cutting wheel turns between closely fitted guides, moving up and down between the guides and across the rods being cut The process is wet cut ting and the wheel is not visible to the operator while cutting Most machines cut two rods at a time

The operator pushes the rods through guides into stops which regulate the thick ness of the discs Holding the rods firmly against the stops with a hand lever, he operates a pedal with the left foot which controls the cutoff wheel as it is applied to the rods As soon as he cuts through the rods he lifts his foot, the wheel rises, and immediately after it has cleared the stops a backward jerk of the right hand actuates ejector or knock out pins which knock the severed discs into the stream of water Immediately after the ejection of the discs the rods are pushed back into the stops and a new cut is taken Figure 11 1 shows an operator seated at a machine ready for operating

The cutoff machine depends for success ful operation upon the speed, form, rhythm, and pressure pattern of the hand and foot action of the operator Failure to apply foot pressure properly results in damage to the discs, excessive breakage and use of wheels, and wastage of material

The purpose of this study was to ana lyze the disc cutting operation by identify ing the form of the foot movement that produced satisfactory quantity and quality of discs with minimum cutting wheel usage, and to teach this form to new operators by the movement analysis method

THE METHOD

The apparatus The mechanical appara tus used for the movement analysis in cluded a paper tape recorder with a specially made writing arm attached A fine thread was fastened to the writing arm and run through glass bushings mounted in metal pieces which were clamped at accessible places on the cutoff machines, and thence to the pedal in such a manner that the complete cutting cycle movement could be recorded on the paper tape as it moved along at a known speed under the pen The recorder was placed on a small wheeled table which was pushed from machine to machine and checks on oper ators were quickly and accurately made without interference with production Fig ure 11 2 is a schematic drawing of the re corder attached to the cutoff machine

Procedure The procedure consisted first of a job analysis by activity The principal

* Reprinted from *Journal of Applied Psy chology* Vol 29, No 6, December 1945

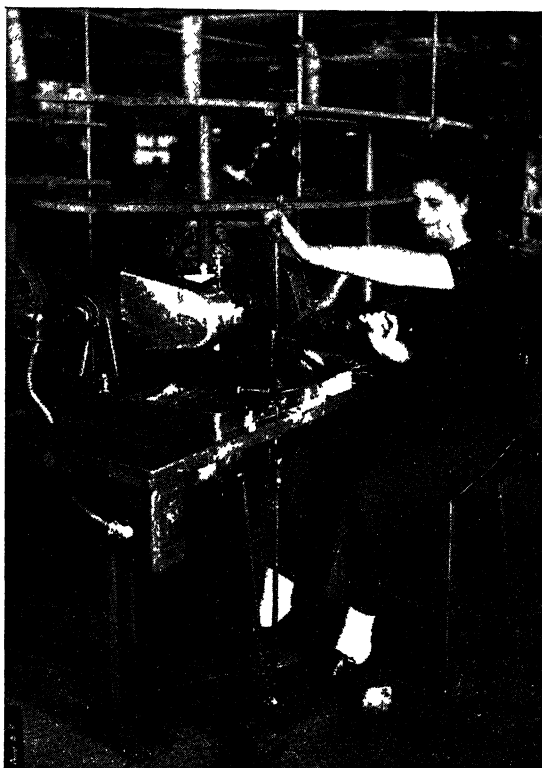



FIGURE 11 1 *An operator seated at the disc cutoff machine ready for operating*

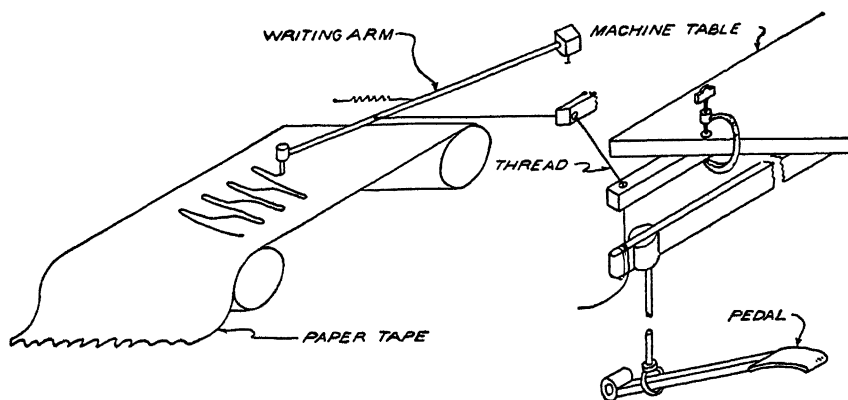



FIGURE 11 2 *Schematic drawing of the paper tape recorder attached to disc cutoff machine*

activity, the cutting cycle, was further analyzed by studying the foot action recordings of skilled operators to find out what constituted good performance. The recorder was then used for instructional purposes in the training of new operators and for improving experienced operators already on the line. This type of analysis is not entirely new. It was utilized with good results by Tiffin and Rogers (5) in training tin plate inspectors and by English (1) in training riflemen.

The coordination of the foot action in cutting, with the hand action in placing the rods against the stops and ejecting the discs after cutting, constituted the cutting cycle. The cut through the rod was the cutting phase and the ejecting of the discs and placing of the rods against the stops was the recovery phase.

Since the foot action was the principal part of the cutting cycle, the main effort was given to finding the correct foot action pattern and then presenting it to the trainees in an effective manner.

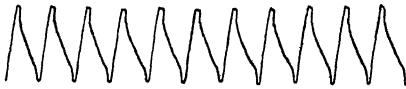



FIGURE 11 3 *Disc cutter foot action pattern of a good experienced operator. This was the accepted standard.*

By recording the patterns of experienced operators and supervisors, who were also experienced cutters, and comparing the action patterns with quantity, quality records, and abrasive wheel usage records, it was possible to identify the "standard" or correct pattern. There were wide variations in action patterns of the experienced operators showing the existence of individual differences, but the consistent pattern was there to form the basis on which the noted differences could be explained. Figure 11 3 shows the foot action pattern accepted as standard. All sizes of rods gave the same pattern, but spacing of cuts was wider for large rods because of longer cutting time.

The correct pattern showed a good reverse cutting curve which indicated unhurried, steady cutting with "feel" of rod. The operator checked his foot at the same place on every cut just as he cut through the rod by easing the pressure, thus allowing the wheel to cut its way through without making burrs. The short follow through at the bottom indicated, in shop parlance, a "soft" foot. The recovery started soon after cutting through, saving time for each cut and allowing plenty of time to coordinate or time the ejection of the discs and place

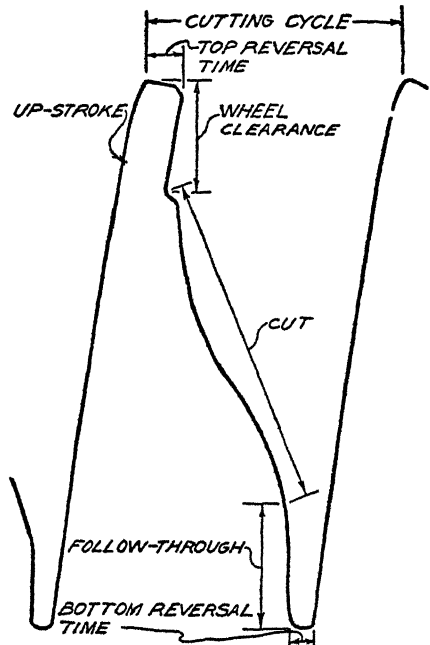



FIGURE 11 4 *An enlarged view of disc cutting cycle showing its principal parts.*

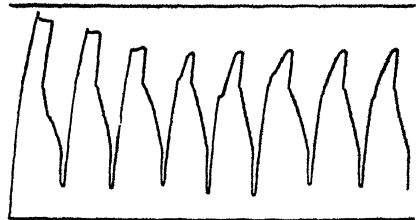



FIGURE 11 5 *An incorrect disc cutter foot action pattern. Operator had only 4 hours experience.*

the rods back against the stops ready for the next cut The operator paced the machine and adjusted his time for the cutting part of the cycle to the speed at which the wheel could cut best without forcing or crowding the wheel To force or crowd the wheel results in several disc defects, short wheel life, and wear out of the setup Figure 11 4 is an enlarged view of one cycle showing the significance of each part of the cycle Figure 11 5 is an incorrect pattern made by a trainee with only 4 hours experience Figure 11 6 shows the foot action patterns of one trainee at various stages in the training program There is a steady approach to the standard pattern and a complete story is evident in each recording after the various hours of supervised operation It will be noted that the pattern at 239 hours closely resembles the standard pattern shown in Figure 11 3

Enlarged instructional posters with analytical notations were prepared both for correct and incorrect patterns and for various types of damage to the product which were shown to be reflected in the action pattern Individual action patterns of each operator were kept in folder form also so the operators could note their progress By careful use of the recorder at timely intervals with both the trainees and experienced operators and by interpreting the foot action patterns in terms of the standard, it was possible to reduce the training time of new operators and to improve the quality and quantity performance of some employees already on the job

Results

Trainees versus old operators Training a group of operators on the production line is not unlike breaking in a team of horses, a group of men for a football team, or a squad for the army All must work together Especially is this true where the operators are working on day rate and a certain amount of production must be obtained each day One individual cannot do all the work no matter how good he is The object of the training program is, therefore, to get all of the trainees as nearly alike as possible in the shortest time and all as good and as near like the standard as possible The supervisor, or trainer, has to move from operator to operator, checking, helping, comparing, encouraging, correcting, explaining, and expecting progress day by day If he does not maintain his own interest he will not find continued interest in the learners, especially after the newness of the job has worn off He should be interested in the group, it is from the group that he expects to obtain production It is, therefore, on the group that the training program must be evaluated Even when all possible attention has been given to the

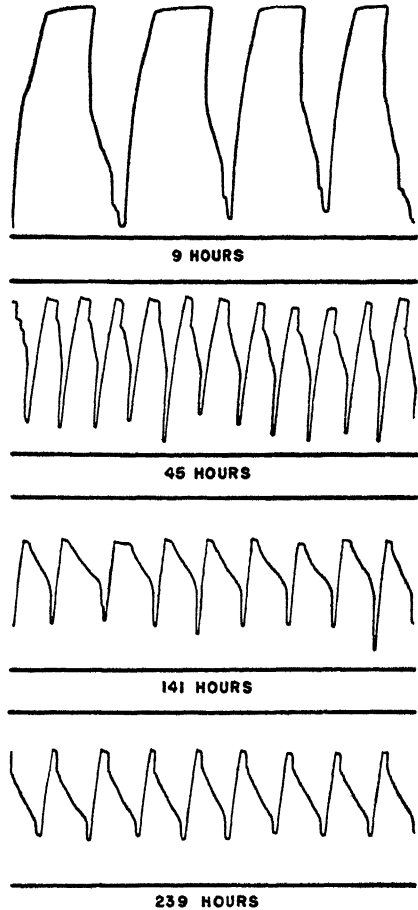



FIGURE 11 6 *Disc Cutter Foot-action Patterns of a Trainee Showing Improvement With Training The records were made after 9 45 141 and 239 Hours of Supervised Operation*

group, individual differences will remain Tiffin (6) enumerates many kinds of individual differences in job qualifications and individual productivity The person doing the training must constantly detect and give due consideration to these individual differences in order to get the entire group to the stage of perfection he desires Equal increments of practice and training do not in any sense reduce all trainees to a common level of performance In evaluating the results of a training method on the job where none of the variables is controlled, there are factors that prevent one from giving too much attention to the production of any one individual For instance, there are external limitations on the individual s power to increase his output, such as the condition and speed of his machine One trainee may be given credit for less output than another because he was carrying an undetected handicap And even though this handicap is detected there is often no way of eliminating it The speed of two machines that look exactly alike may vary tremendously during different periods of the day, or supplies of material or the difficulty of the task may set an upper limit on individual production which perhaps one-third of the members of the group could reach, and no one could better, no matter how competent he might be Age is another factor which has unmeasured effects on individual production on the line and makes it necessary to consider group differences in evaluating the results of the training program

There are numerous ways to evaluate the results of the training once it has been put into operation Lawshe (2) lists 13 methods and from such an extensive list one should find a method that fits any particular case

In machine operations, like disc cutting, jobs have been time studied, rates based on these studies have been set, and production of trainees and old workers alike is indicated by the output In the particular operation reported herein training can be evaluated on two factors, namely, production performance and wheel performance Both wheel breakage and wheel use are important aspects of wheel performance The two factors are considered to be of equal importance and management has set up a rate sheet based on the number of cuts secured per wheel for a given size of rod and the speed of cutting (production per hour) In other words, the operator must get a specified number of cuts per wheel and must cut at such a speed that he will turn out enough units in one hour to earn the base rate set per hour When he gets the exact number of cuts per wheel for the rod size he is cutting and cuts just enough to earn his hourly rate, he is doing 100 per cent performance in both factors, namely, production performance and wheel performance

There are several reasons for emphasizing wheel performance One is that the special type of wheel used for tungsten disc cutting is very expensive and wheel costs can very easily exceed labor costs Also, a better quality of product is obtained by limiting the speed of cutting, which is effectively done by placing emphasis on wheel performance

Curves are used to evaluate the effects of this training program As stated by Tiffin (6, 186), ' such curves, therefore, serve the very useful purpose of providing a means of evaluating the successfulness of an operator training program and spotting decisively those operations in which training is inadequate, either in quantity or quality

Figure 11 7 is the production curve of the trainees for 12 weeks Production perform

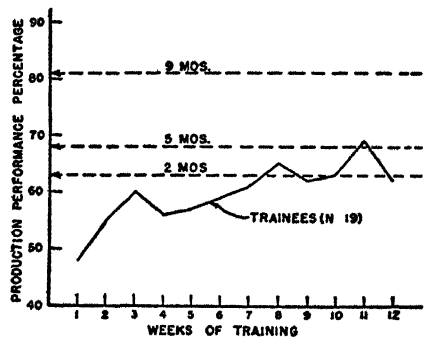



FIGURE 11 7 *Production Performance percentage of disc cutter trainees Dotted lines represent average production performance percentage for first 2 weeks of groups with an average of 2 5 and 9 months experience at start of training*

ance percentage is plotted on the vertical axis and weeks (6 days each) of training on the horizontal axis Production perform ance percentage makes the production per formance of all operators comparable since it is computed on the company rate sheet for the various sized rods cut For instance, if an operator using 6 wheels to cut 142 inch rods, cuts 12,000 units of rod in 10 hours, he cuts at the rate of 1,200 units per hour The rate sheet indicates the op erator should cut 2,340 units per hour to earn the base rate of 65 cents per hour Since 1,200 is 51 per cent of 2 340, his production performance is 51 per cent The production performance percentage is, therefore, the average number of units cut per hour divided by the number of units that the operator is expected to cut per hour of the size rod being cut to earn his wage

The dotted lines in Figure 11 7 show the production performance of old operators at the start of the training experiment There are three groups of these old oper ators Before the training started, eight operators had an average of 2 months, nine operators had an average of 5 months, and ten operators 9 months experience on the job The trainees had no previous expe rience Old operators total 27 and trainees 19 No data were available on the produc tion abilities of the various groups prior to the start of training In order to obtain a basis of comparison for the groups, an av erage of the production of the first two weeks at the beginning of the training is considered indicative of the ability of the groups of old operators at that time These averages are shown on the figure by the dotted lines

The production curve for the trainees reflects a steady rise At the end of 8 weeks they have surpassed the average of the 2 months group At 11 weeks they exceeded what the 5 months group was able to do at the beginning of the training program Although their average fell on the 12th week, this is not unusual in curves of this type The general trend is upward and with more confidence, experience, and con tinued training it should be safe to assume that the curve should reach the top limit of the 9 months group in much less time than 9 months, resulting in saving many man hours of training time

One may legitimately inquire about the quality of the discs cut There was no dif ference in the quality of the discs cut by the trainees and by the other groups The company used a production checker who continually checked the discs as they were cut by going from operator to operator and it was practically impossible for any one operator to cut a majority of the de fective discs The machine would be shut down or the operator s performance inves- tigated to determine the cause of the de fects Any shut down would be reflected in the production performance of the par ticular operator Since tungsten is an ex pensive material, it was imperative to keep wastage at the very minimum One hun dred per cent inspection of the discs was made after a tumbling operation, but this was done according to size of discs and the production of two or three operators cut ting the same size rods was usually thrown together Production control regulated the quantity to be cut over any one period In the form of stricter supervision this con trol probably served as a forced incentive to the worker

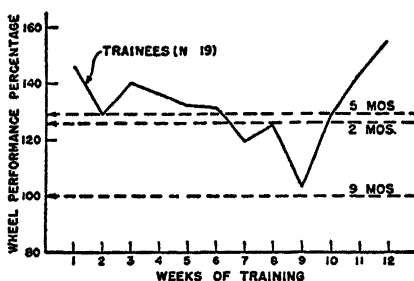



FIGURE 11 8 *Wheel performance per centage of disc cutter trainees Dotted lines represent average wheel perform ance percentage for first 2 weeks of groups with an average of 2 5 and 9 months experience at start of tra ning*

Figure 11 8 is the wheel performance per centage plotted against weeks of training The averages of the 2, 5, and 9 months groups of old operators in wheel perform ance are shown by the dotted lines These averages are of the first two weeks per formance at the beginning of the program

Wheel performance percentage is computed from the rate sheet. Referring to the illustration previously cited, if the operator who cut 12,000 discs in 10 hours used, without breaking, 20 of the 6″ wheels and broke 5 others, he used a total of 25 wheels to cut 12,000 discs. He obtained 480 units per wheel. According to the rate sheet, 100 per cent wheel performance on .142 rod requires 470 units per wheel. The wheel performance is 480 divided by 470 or 104 per cent. Wheel performance is, therefore, the units the operator gets from each wheel divided by the number he should obtain according to the rate sheet for the size of rod being cut. Wheels broken must be counted as wheels used, for the operator gets some cuts with the wheels before they are broken. Wheel performance is negatively correlated with production performance. The faster the cutting the higher the production and the fewer units per wheel. This relationship had to be regulated in the training program and emphasis placed on cutting according to the cutting curve so that both maximum wheel performance and maximum production were obtained.

The rate sheet is made up on the basis of 100 per cent production performance and 100 per cent wheel performance. Figure 11.8 shows the 9 months operators at exactly 100 per cent wheel performance while Figure 11.7 shows their production performance at 81 per cent. It will be observed by reference to Figures 11.7 and 11.8 that the trainee curve goes up toward 100 per cent production performance and downward toward 100 per cent wheel performance.

One must speculate on the sudden upturn in the wheel performance curve of the trainees after the 9th week. It can be accounted for to some extent at least by the fact that a more rigid accounting of wheels used was started at that time. Some operators had acquired the habit of not using all the wheel because of the slower cutting pace due to the smaller diameter of the wheel as it wore down. These operators changed wheels before entirely using them. This lowered the wheel performance and took more wheels. By requiring operators to turn in every used wheel for inspection at the end of the day's run it was possible to determine just which operators had not entirely used their wheels. The general effect was that all operators tended to use up their wheels, thus getting higher wheel performance. This might also cause the slightly lower production because of slower cutting as the wheel grew smaller.

The best operator would get as much over 100 per cent wheel performance and as much over 100 per cent production performance as possible. To do this he would have to cut steadily, consistently, and carefully during all working time. Failing to do this he could not possibly get over 100 per cent wheel performance even though he might cut fast to make up lost production time, because the faster he would cut the lower his wheel performance would become. The interrelationship of wheel performance to production performance complicated the training problem and made the movement analysis method all the more important. As has been noted, this found a standard pattern that laid stress on both accuracy and wheel performance.

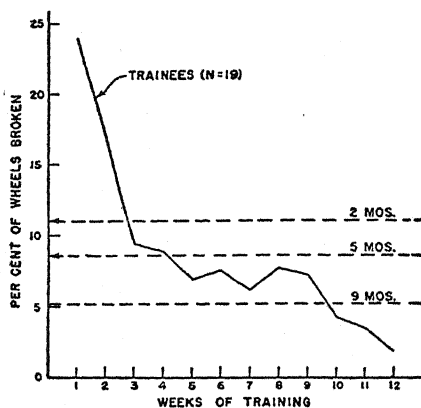



FIGURE 11.9. *Percentage of wheels broken by disc cutter trainees. Dotted lines represent average percentage of wheels broken for first 2 weeks by groups with an average of 2, 5, and 9 months experience at start of training.*

Figure 11.9 shows the percentage of wheels broken plotted against weeks of time. Percentage of wheels broken is the fraction of the total number of wheels used

each week For instance, the number of wheels used without breaking plus the number broken is the total number used The ratio of the number broken to this total is the percentage broken If, for example, an operator wore out 300 wheels in one week and broke an additional 25, he used a total of 325 wheels The percentage of wheels broken during the week is 25 divided by 325, or 7 7

Figure 11 9 shows graphically the trainees age In Figures 11 7, 11 8, and 11 9 the reduction in wheel breakage due to the attention given to the correct cutting method During the first week the trainees broke 24 per cent of their wheels By the third week they were breaking less than operators having 2 months experience, by the 5th week less than those having 5 months experience, and by the 10th week less than those having 9 months experience By the 12th week of training the trainees broke only 1 8 per cent of their wheels, or about one third that of the group with 9 months previous experience This is good for two reasons, first, because the wheels were very expensive and large savings resulted (wheel cost could very easily exceed labor cost), and second, all time taken to change broken wheels has to be charged against production time Broken wheels prevent running for a period ranging from one minute to sometimes as long as a half hour, depending on how long it takes to remove the broken pieces from between the guides

The higher average wheel performance noted for the trainees is due to emphasis being placed on accuracy rather than speed and on correct operation, or cutting according to the reverse cutting curve High wheel performance eventually leads to increased production It insures a better quality of product at the beginning with less wastage The curve shows that the trainees steadily reduced their wheel break trainees show exceptionally fine performance in relationship to the other three groups of operators who learned bv the "pick up" method Although their production average was lower than the 9 months group of operators, they steadily increased production, their wheel performance was maintained well over 100 per cent, and the per cent of wheels broken was considerably reduced

Trainees versus beginners Another way operators learn jobs, probably more prevalent than any other learning method in industry, is that of ' trial and success This consists of showing the operator his job and then ' turning him loose to learn in the best way he can It is a sort of sink or swim proposition The operator is not led from the beginning by progressive steps and has no reliable indication of his progress Such haphazard learning usually results in lower efficiency and poor work habits

It was possible to obtain records over a period of three weeks of the production

(*a*)

(*b*)

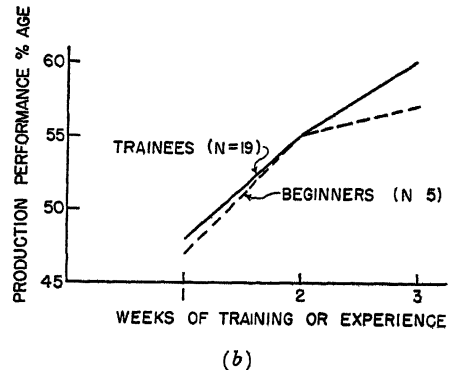
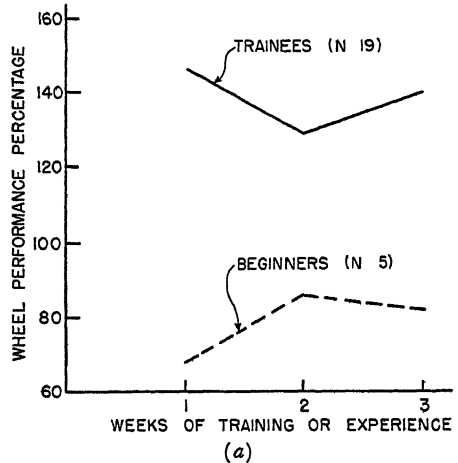



FIGURE 11 10 *Wheel performance and production performance percentage of disc cutter trainees and beginners for first three weeks*

and the wheel performance of 5 operators who were learning the disc cutting operation in just this manner They are called beginners for purposes of comparison with the trainees who were trained by an organized method These beginners did not have the advantage of using the movement analysis recorder nor any of the instructional material

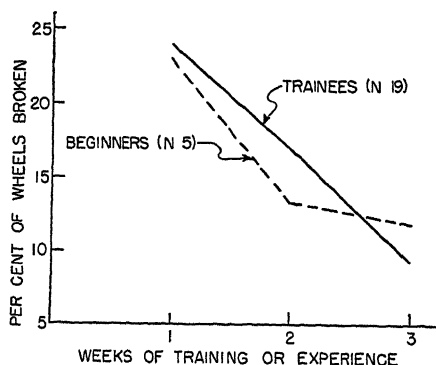



FIGURE 11 11 *Percentage of wheels broken by disc cutter trainees and beginners for first three weeks*

Figures 11 10 and 11 11 present a comparison of the production performance, wheel performance, and wheel breakage of beginners and trainees at the end of the first, second, and third week of operation on the job Figures 11 10(b) and 11 11 do not show a significant difference between the two groups in production performance and percentage of wheels broken Figure 11 10(a), however, shows graphically considerable difference in wheel performance percentage between the two groups Whether the difference shown can be attributed to chance can be tested statistically by finding (t) the significance ratio for the difference in the means of two independent small random samples The smallest difference between the two curves occurred at the second week The significance of this difference, therefore, was computed to determine the possible role of chance in accounting for the difference between the curves The second week of training or experience shows a wheel performance percentage mean of 129 for the trainees and 86 for the beginners Substitution in the formula (3) for finding the significance ratio for the difference in the means of two independent small random samples results in a 't" of 2 548 Reference to the table of t's indicates a "t' of 2 518 for 21 degrees of freedom is required for the 2 per cent confidence level The obtained difference is therefore significant at the 2 per cent level which means that there are 98 chances in 100 that the difference found is not due to chance and must therefore be due to the special training given the experimental group

Effect of training program on old operators When any training is carried out directly on the production line it is bound to have some effect on the experienced operators even though most of the attention is directed to the trainees Studies have been made of factory workers in which work conditions of all kinds were varied A noteworthy example is the experiment carried out in the Hawthorne plant of the Western Electric Company (4) Its most important finding was that regardless of the nature of the changes made in the working conditions the productivity of the experimental group tended in general to increase Although the interest and attention undoubtedly had a stimulating effect on the old operators, the fact that foot action recordings and interpretations were made for them as well as the trainees would lead one to believe that interest and attention did not account for the total improvement

One cannot rush up to an old operator with a new gadget with the expressed purpose of checking up on him The cooperation of the old operator may be secured, however, by explaining to him the need for finding out how the job is done so that it can be taught to the new operators Practically every old operator will give the best performance he can muster up and is then already on the road to self improvement One can also question the operator about his work As soon as the old operator finds that someone is interested in his job he begins to think that the job is important after all and he is likely to take more interest in his work and that of fellow workers Learning becomes contagious, production starts to increase, the entire

lme ımproves, and no one but the person ın charge of the traınıng ıs aware of ıt or knows the reason why

If those operators who seem not to want help are ıgnored for a whıle they wıll soon seek help eıther by askıng for a recordıng or by lettıng ıt be known that theır record ıngs have not been taken Workers all de sıre to be treated alıke

A revıew of the foot actıon patterns[1] shows that old operators used a lot of waste motıon ın cuttıng and most of them had not developed the necessary feel That they could ımprove by watchıng the re- corder was also shown

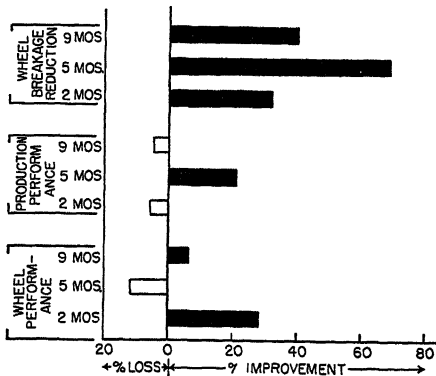



FIGURE 11 12 *Percentage of ımprovement or loss ın wheel performance productıon performance and wheel breakage reductıon for groups havıng an average of 2 5 and 9 months experıence on the dısc cutoff machıne prıor to the start of the traınıng program*

Fıgure 11 12 shows the percentage of ım provement or loss ın wheel performance, productıon performance, and wheel break age reductıon made durıng traınıng by groups havıng an average of 2, 5, and 9 months of experıence on the dısc cutoff machıne prıor to the start of the program The graph ıs based on averages for the first three weeks and the last three weeks of the traınıng perıod In computıng the percent age of ımprovement or loss, the perform ance at the start of the program was consıdered 100 per cent

All three of the groups broke fewer wheels after the traınıng was under way The 5 months group made the greatest ımprovement by breakıng 69 per cent fewer wheels than they dıd at the start, the 9 months group 40 per cent less and the 2 months group 32 per cent less Reductıon ın wheel breakage effected a substantıal savıng ın wheel costs over the perıod

The 5 months group ımproved theır productıon performance 21 per cent but suffered a loss of 11 per cent ın wheel per formance Sınce productıon performance ıs negatıvely correlated wıth wheel per formance and both are of equal ımpor tance, thıs represents total ımprovement and reflects the emphasıs placed on wheel performance by the movement analysıs method Lıkewıse the 2 and 9 months groups ımproved theır wheel performance by 28 and 6 per cent, respectıvely, at the expense of productıon by only 6 and 5 per cent Agaın wıth wheel performance and productıon performance of equal ımpor- tance thıs represents substantıal ımprove- ment

The total pıcture ıs good when one con- sıders the nature of the operatıon, that ıt ıs a hand and foot coordınated operatıon whıch requıres the human operator to functıon wıth mechanıcal precısıon over long perıods of tıme, and that old opera- tors may have learned as many bad habıts as good, necessıtatıng unlearnıng before startıng new habıts

That these experıenced operators could so effectıvely cut down ın the number of wheels broken and ımprove ın both wheel and productıon performance at thıs stage of experıence ındıcates that learnıng of a better cuttıng cycle by old, experıenced operators took place wıth the ınstallatıon of the program for traınıng new operators The ımprovement made by the experıenced operators as well as the traınees seems to justıfy movement analysıs as an effectıve ındustrıal traınıng method

SUMMARY AND CONCLUSIONS

The problem of traınıng rapıdly a num ber of new workers to operate certaın types

[1] Case studıes and complete presentatıon of data wıll be found ın the appendıx of a thesıs by the author on file ın the lıbrary of Purdue Unıversıty

of cutoff machines used for cutting tung sten rods into small discs for electrical ap paratus has been explained A method has been described which analyzes the foot movement of the operator by the record ing of a pattern on a moving paper tape A standard pattern was established and the recorder was used to take recordings of the trainees at frequent intervals These recordings were explained on the basis of the standard Various defects in the prod uct were identified with the foot movement patterns and many individual differences in both old and new operators were noted and described as clinical case studies Finally, complete production records of all operators were kept These records were plotted in graphical form to show the progress of the various groups at different stages of training

In general, the findings warrant the following conclusions

1 The disc cutting operation was analyzed by activity and the best form of cutting movement identified

2 The cutting movement of the oper ators was shown in graphic form, compared with a standard, and used as an effective training method

3 Training time was effectively reduced At 8 weeks the production performance percentage of the trainees was better than that of old operators who had had the same average amount of experience by the 'pick up' method

4 Trainees, who were taught to use the cutting wheel by studying the movement analysis recordings, secured better wheel performance than old operators

5 Trainees reduced their wheel break age in a few weeks to a point much lower than the average of old operators at start of training program

6 Trainees did better in wheel perform ance than beginners who learned by the 'pick up' method

7 Experienced operators benefited by the training program

8 It is entirely possible that equally suc cessful improvements in training could be made in numerous other manipulative jobs by the application of a similar method of recording graphically the performance of the operator

REFERENCES

1 English, H G, How Psychology Can Facilitate Military Training, *Journal of Applied Psychology* 1942, Vol 26 3–7

2 Lawshe, C H, Jr "Training Opera tive Personnel *Journal of Consulting Psychology*, 1944, Vol 8 158

3 Lindquist, E F, *A First Course in Statistics* New York Houghton Mifflin Company, 1942 pp 138–139, 240

4 Mayo E, *The Human Problems of an Industrial Civilization* New York The Macmillan Company, 1933, Chs 3–5

5 Tiffin, J, and Rogers, H B, 'The Selection and Training of Inspectors,' *Personnel* 1943, Vol 22, 3–20

6 Tiffin, J, *Industrial Psychology* New York Prentice Hall, Inc 1942, pp 1–19

Testing a Training Program in Human Relations *

RAYMOND A KATZELL

IN BRIEF

This study was performed to evaluate a training program designed to improve understanding of human relations on the part of a group of experienced supervisors The program consisted of a series of con ferences conducted by the determinate dis cussion method and devoted to various topics drawn from modern industrial psy chology By way of evaluation, alternate forms of the File Remmers "How Supervise?" questionnaire were administered to ward the beginning and end of the

* Reprinted from *Personnel Psychology*, Vol 1, No 3, Autumn 1948

program This questionnaire calls for expressions of attitudes toward supervisory principles and practices The average score of the trainees was significantly higher toward the conclusion of the program than it had been toward the start Improvement was found to be greater on the part of those trainees who had comparatively low scores initially, who had relatively less supervisory experience, and who obtained higher scores on a test of intellectual alertness As another aspect of evaluation, the former trainees were requested to rate the program anonymously as to its interest, usefulness, and over all value The vast majority of ratings were very favorable The evidence suggests that the program was effective in improving understanding of the human problems in supervision and that it may also have had some value as a morale builder among the supervisory staff

Discussion

Industry now recognizes that supervisors and foremen are among the most important girders in the complex structure of employee management relations In their efforts to strengthen this structure, a growing number of companies are conducting programs for training present and prospective supervisors in sound principles of human relations

The personnel official who desires to initiate such a program, or to improve his present one, naturally asks what is the best kind of program for his needs Unfortunately, there is not enough evaluative information available on the basis of which his question can be readily answered What is needed is a large number of reports of training programs in which the following information is included purpose of the program, types of personnel involved, method by which training was conducted, topics included, and results of the program When a body of such data has been amassed, it should be possible to eliminate much of the guesswork involved in designing a suitable program

The following report of a human relations training program[1] is submitted as a contribution to this body of information

Purpose of the Program

The program was designed to develop in experienced supervisors a more thorough comprehension of sound principles of human nature and interpersonal relations It was hoped that this would lead to greater effectiveness in supervision of subordinates and in dealings with superiors and coordinates

The undertaking was designed as a job improvement program, and should therefore not be viewed as a complete training course such as might be desired for prospective or inexperienced supervisors

Who Were Trained

The trainees included in the present program were employed in supervisory positions in two divisions of a large railroad They were drawn from nearly all departments of their organization, the basis of selection being their classification in intermediate grade supervisory positions (i e, above foreman but below executive grade) In the period covered by this report, seventy three men started in the program However, complete evaluative data were obtained for only sixty of them, and it is on these men that the results to be described are based

The trainees were for the most part highly experienced, having had an average of 18 6 years of experience in supervisory work as foremen or higher On the average, they had completed 10 5 years of schooling

[1] This was one of a series of such programs sponsored by the Illinois Central Railroad in its various divisions The training was given by different individuals in the various states in which the railroad operates, and the programs therefore differed somewhat from region to region The first program was organized by the University of Illinois for the divisions in that state This report is of the program conducted by the author during June and July, 1947, while at the University of Tennessee, for the railroad s divisions in that state The entire program series was under the supervision of Mr C R Young, Director of Personnel of the I C

How They Were Trained

The men in each of the two divisions were divided into two groups, the largest of the four groups having twenty two members and the smallest, sixteen Each group met one day a week for about two hours in the morning and two in the afternoon Meetings were held during eight consecutive weeks for a total of thirty two hours

The meetings were held on company time, although many of the men were on twenty four hour call Attendance was required Positive incentives included emphasis on the opportunity for self improvement and the awarding of a non credit certificate from the University of Tennessee upon completion of training No grades or examinations were given

The meetings were devoted to conferences conducted by the determinate discussion method Beckman (1) has described five salient features of the method to be the following (a) each discussion meeting has specific objectives, (b) the discussion is directed by the leader to conform to a predetermined outline, (c) chart summaries are prepared in advance to serve as a guide for the leader, who endeavors to have the group suggest the more important points—duplicate copies of these charts are later distributed to participants to serve as reference notes, (d) emphasis is laid on citation of specific case material by the group, (e) the leader takes an active part in directing the path of discussion, but does not attempt to force his own opinions on the group

Sections of the book *New Techniques for Supervisors and Foremen* by Walton (6) were recommended to be read in conjunction with the various conference topics

Topics for Discussion

The general point of view and the topics selected for discussion were derived mainly from modern industrial psychology However, the discussions were held on a non technical plane, with emphasis on basic principles and their practical applications to human relations in industry In planning the topics and their details the author was influenced considerably by his experience with earlier groups of trainees in the same company

Approximately two hours of conference time were devoted to each of the following topics, which were grouped under three general headings

The Supervisor and His Job

Topic A Introduction and orientation
Topic B Responsibilities of supervisors and management
Topic C Characteristics of the top notch supervisor

The Supervisor and Human Nature

Topic D Human nature and the industrial scene (The scientific approach to human nature, factors affecting performance of industrial workers)
Topic E How and why people differ from one another in ability
Topic F Efficient conditions of working
Topic G Developing skills I
Topic H Developing skills II
Topic I Understanding and developing personality
Topic J Understanding peculiar personalities

The Supervisor and Leadership

Topic K What men want out of life and work
Topic L Motivating and getting cooperation from others
Topic M Leadership and the development of morale
Topic N How to lead and deal with people
Topic O Accidents a challenge to leadership
Topic P Making yourself a more effective person

Testing the Training

Ideally, a training program of this sort should be evaluated in terms of the on the job behavior of those with whom the trainees come in contact One should like to know what changes occurred over a period of time in the productivity of the workers, in rates of accidents, absenteeism and turnover in grievances, in suggestions and the like But such evaluations require preparation for the acquisition of data well

in advance of the actual program, as well as provisions for follow up after the train ing program In short, such evaluation ordinarily requires a rather extended period of association both prior to and fol lowing the training program, together with specific planning for evaluation It is heartily urged that this be done, where feasible

However, in the present situation these requirements could not be met and the program had to be evaluated, if at all, by more immediately obtainable criteria It was decided to do this in terms of expres sions of attitudes and opinions of the train ees (a) toward supervisory policies and practices and (b) toward the training course itself

Changes in Supervisory Attitudes

In order to evaluate what influence the program had on points of view toward practical supervisory problems, the File Remmers How Supervise?' questionnaire (2) was administered at the start of the second conference, and an alternate form of the same questionnaire was used at the end of the fifteenth conference in the series Form B was the one administered toward the beginning and Form A toward the end These forms will hereafter be referred to as the initial and terminal forms, respec tively Although the two differ in the spe cific content of items they cover the same general areas of opinion and have been equated in over all difficulty Each form contains seventy statements, seventeen on supervisory practices, twenty four on com pany policies affecting personnel, and twenty nine on supervisory opinions on human factors in industry The person an swering the questionnaire is instructed to indicate whether he endorses or disagrees with each one of the statements, or is un decided in his opinion He is given one point of credit for each statement on which his judgment corresponds with that of a majority of personnel and training special ists who answered the questionnaire during its developmental stage

The trainees were informed that the questionnaires would not be used to eval uate any individual nor would any man s answers be divulged

When the questionnaires were scored after the conclusion of the training pro gram, it was found that the average score for the entire group of sixty trainees was significantly higher on the terminal ques tionnaire (43 3) than on the initial one (37 8) This trend held also for each of the two company divisions considered separately What was found in short, was that as a result of the training program the judgments of our trainees on super visory practices and principles changed so that they became more like those of a group of personnel and training specialists

The amount of improvement is indicated by the fact that the average score of our trainees on the terminal questionnaire ex ceeded the scores of about 75 per cent of 557 supervisors in various industries tested by the authors of the questionnaire, whereas on the initial questionnaire the average performance of our trainees had exceeded the scores of only about 62 per cent of those supervisors

It should be noted that the specific issues posed in the questionnaire were not dis- cussed as such in the conferences It thus seems reasonable to believe that the im- provement was the result of application of more effective general principles to the problems raised, rather than of the learn- ing of specific answers to specific situations

It is possible, of course, that the im- provement was due not primarily to the training but to the 7 weeks' additional ex- perience the men had had as supervisors To check on this point a control group of equally experienced supervisors, engaged in the same kind of work as the trainees but not enrolled in the training program, should have been given the two forms of the questionnaire at the same time as the trainees That group s improvement, if any, could have then been compared with that of the trainees But, unfortunately, there were not enough employees who could qualify for membership in the control group, practically all intermediate level supervisors in the two divisions were being or had already been trained Nevertheless it seems unlikely that much of the improve ment noted in the trainees was due to the

job experience accrued between the administration of the two questionnaires It will be recalled that the average supervisory experience of the group at the inception of training was over 18 years, and it is not likely that an additional 7 weeks would have made much difference

In order to shed light on some of the factors that might be related to the effectiveness of the training program, the relationships between improvement on the questionnaire and certain personal characteristics of the trainees were studied Analysis of the data brought out three pertinent facts

1 There was a definite inverse relationship between improvement on the questionnaire and effectiveness of attitudes at the inception of training Those men whose scores on the initial questionnaire were comparatively low tended to improve more than those who had high scores initially Of course, the possibility exists that it is harder to improve performance on the questionnaire once one reaches its upper range and that this ceiling effect accounts for the foregoing finding But this possibility remains in the realm of conjecture as far as this study is concerned

2 When the influence of differences in scores on the initial questionnaire was statistically eliminated those trainees who improved most tended to be those who received the higher scores on a test of intellectual alertness (5) which had been administered to the entire group

3 There was a definite tendency for the less experienced supervisors to improve more than the more experienced ones The data indicate that this cannot be attributed either to less knowledge at the beginning of the course on the part of the newer men or to greater mental alertness Possibly the newer men, being less set in their ways, found the conclusions reached in the course less in conflict with habits of long standing

From these results it may be deduced that the training program was most effective for supervisors whose opinions were most different from those of the experts to begin with, who were not highly experienced, and who were relatively 'bright' These deductions, of course, apply when effectiveness of the training is evaluated in terms of improvement on the How Supervise? questionnaire

How the Trainees Rated the Course

One of the grounds for judging the adequacy of a training program should be the evaluation of the program by the trainees who participated in it One would certainly be justified in wondering how a group of experienced, practical railroad supervisors, having on the whole relatively little formal schooling, would react to a training course of the type described here To check on this matter the following procedure was employed

Six months after the conclusion of the course a rating form was distributed to each of the former trainees On this form the men were instructed to rate the program in regard to three aspects interest, usefulness in connection with their work, and over all value to them Each aspect was rated by checking the one of five categories which best described the rater's judgment of the course The men were urged to be frank in their appraisals The completed rating forms were returned anonymously through the mails Usable ratings were returned by fifty seven of the trainees

In regard to how interesting the program was to them, 93 per cent of the raters judged the program in one of the two highest categories ("very interesting' or 'interesting") Eighty two per cent rated the program in one of the two highest categories with respect to usefulness ("very useful' or "useful") Over 89 per cent rated the program in one of the two highest categories relative to over all value ("very worthwhile or "worthwhile")

The appreciation shown for the program not only supports the aforementioned evidence that it was an effective training procedure, but also suggests its possible value as a morale builder among the supervisory staff

Technical Section

This section is devoted to the statistical treatment of the data on which the results reported in the preceding sections are

based Those statistics marked (*) are significant at the 01 level, those marked (**) are significant between the 01 and 05 levels

A Comparison of average performance on initial and terminal questionnaires for entire group of trainees (N = 60)

	Initial (Form B)	*Terminal (Form A)*
Mean	37 8	43 3
σ	13 6	12 7

$$\frac{M_A - M_B}{\sigma_{diff}} = 3\ 98^*$$

B The same analysis is shown below for each division separately

M Division (N = 32)

	Initial (Form B)	*Terminal (Form A)*
Mean	35 4	41 5
σ	13 9	12 2

$$\frac{M_A - M_B}{\sigma_{diff}} = 3\ 55^*$$

J Division (N = 28)

	Initial (Form B)	*Terminal (Form A)*
Mean	40 7	45 4
σ	13 0	13 1

$$\frac{M_A - M_B}{\sigma_{diff}} = 2\ 13^{**}$$

The difference between the mean gains of the two divisions is not statistically significant $\left(\frac{Diff}{\sigma_{diff}} = 0\ 52\right)$

The remaining analyses were performed on the trainees of both divisions combined

C The product moment correlation was computed between scores on the initial questionnaire and improvement from the initial to the terminal questionnaire This r was corrected by Thomson's method (4) for the negative influence of unreliability in the instrument by which gains were measured, *viz* the questionnaire In this correction, the reliability coefficient of 76 for each form of the questionnaire, as reported by File and Remmers (2), was used, also used was the computed correlation of 70 between the initial and terminal scores The corrected correlation coefficient between initial status and improvement equals − 36*

D The product moment correlation between improvement on the questionnaire and score on the Adaptability Test was computed, using the Peters and Van Voorhis (3) correction for unreliability in the instrument by which improvement was measured The corrected r = 09, which is not significantly greater than zero In view of the negative correlation between initial status and improvement and the positive correlation between initial status and Adaptability score (r = 50), the partial correlation was calculated between improvement and Adaptability score with the effects of initial status held constant This partial r = 33**

E The product moment correlation was

computed between years of supervisory experience and improvement on the questionnaire The r, corrected by the Peters and Van Voorhis method for the effects of unreliability of the questionnaire, is − 45* (Correlations between experience and initial status and between experience and Adaptability Test score were not statistically significant, the r s being − 06 and − 24, respectively)

REFERENCES

1 Beckman R O, *How to Train Supervisors* (Second edition) New York Harper and Bros 1942
2 File, Q W, and Remmers, H H, *How Supervise?* New York The Psychological Corp, 1943
3 Peters, C C, and Van Voorhis, W R, *Statistical Procedures and Their Mathematical Bases* New York McGraw Hill, 1940
4 Thomson, G H, An Alternative Formula for the True Correlation of Initial Values with Gains,' *Journal of Experimental Psychology* 1925 Vol 8, 323–324
5 Tiffin J, and Lawshe, C H, *Adaptability Test* Chicago Science Research Associates, 1943
6 Walton A *The New Techniques for Supervisors and Foremen* New York McGraw Hill, 1940

Cutting Training Waste *

WILLIAM MCGEHEE

SUMMARY

Many positions in industry require extensive periods of training In one such job in which new employees reached acceptable production rates after periods of training ranging from 6 to 15 months, study of fast and slow learners showed that these two groups could be differentiated as early as the end of the second week of training Separations or transfers at this early stage could be made with 20 per cent better than chance accuracy By the end of the sixth week of training, this advantage over chance had risen to 63 per cent No further increase in accuracy is possible if the results from one group are to be used as a standard for action on another

THE PROBLEM

More and more industrial training men and supervisors have adopted some type of chart, usually based on the individual s production, for plotting the new employee's job progress The curves are used for discussing the learner's progress from period to period, comparing him with other beginners at the same stage of training, for retaining or dismissing him

Employee progress records based on worker productivity, if properly analyzed, can be of definite assistance in training industrial workers In some instances, the use of such records is based on careful mathematical analysis Too frequently, however, they have become part of the training process simply because the user knows of some other company or department using similar methods

When the relationship of early performance on a job to some ultimate level of achievement can be established, the curves are of definite assistance to the supervisor in considering whether or not to retain an employee in training It is the purpose of this study to present a method of determining the relationship between early performance and the time required to reach a standard of acceptable job performance

THE METHOD

There is an operation in the manufacture of rugs involving preparation of spools for

* Reprinted from *Personnel Psychology* Vol 1, No 3, Autumn 1948

the loom which requires a relatively long training period New employees show wide variability in the time required to reach an acceptable standard of production This acceptable standard has been based on time study in the particular mill in volved in this study This standard of per formance is called 'Average Production"

Twenty one employees were trained in this mill during 1946–47 All of them re mained on the job for at least 15 months These operators varied from 6 to 14 months in the time required to reach the standard The average time required was 10 8 months, 9 of the 21 learners reached the standard in less than that time while 12 required longer The 9 operators who reached standard production in less than 10 8 months will be referred to as fast learners, the remaining 12 will be desig nated slow learners The 9 fast learners required an average of 8 4 months to reach standard, the remaining 12 trainees aver aged 12 7 months

The problem now arises as to whether or not the initial job performance of the fast learners was different from that of the slow learners In other words, how early in training could a supervisor determine whether or not a given operator would re quire a relatively short or long period to reach average production? Further, can the performance of these two groups be used as a standard against which to measure the performance of future trainees? In order to secure answers to these questions the average hourly production for each of the 21 operators was computed for each 40 hour work period during the first 8 weeks of on the job training

From the data in Table 1 the average hourly production for the fast group and the slow group was computed for each of the 40 hour periods These averages are shown in Table 13 2 and are graphically portrayed in Figure 13 1

Analysis of Figure 13 1 indicates clearly that the fast learners not only reach stand

TABLE 13 1

Average Hourly Production at Each of First Eight Weeks of Job Training and Months Required to Reach Average Production Each Subject

Subject No	*40 Hour Periods*								*Months to Average Produc- tion*
	1–40	2–80	3–120	4–160	5–200	6–240	7–280	8–320	
1	21	31	27	30	37	42	41	41	6 00
2	24	35	37	53	53	62	65	79	6 00
3	24	35	36	41	53	53	61	48	6 00
4	28	35	39	47	44	53	59	53	6 50
5	09	17	31	35	30	31	39	39	9 00
6	19	?0	22	27	35	41	49	41	9 50
7	14	33	27	26	33	38	39	42	10 50
8	12	20	24	28	28	28	39	60	10 75
9	15	24	24	27	35	26	34	37	10 75
10	13	16	23	21	24	26	31	30	11 50
11	13	11	11	18	19	26	28	36	11 50
12	18	21	22	28	37	34	40	41	11 50
13	17	24	22	23	34	39	37	39	11 75
14	09	13	14	19	22	22	25	31	12 00
15	16	21	23	24	26	31	31	32	12 50
16	10	20	15	15	21	20	20	29	13 00
17	19	22	27	31	29	31	35	37	13 00
18	13	18	24	19	24	40	22	13	13 50
19	12	22	32	32	27	40	40	51	14 00
20	08	17	22	22	31	35	36	28	14 00
21	11	13	17	25	24	28	34	41	14 00

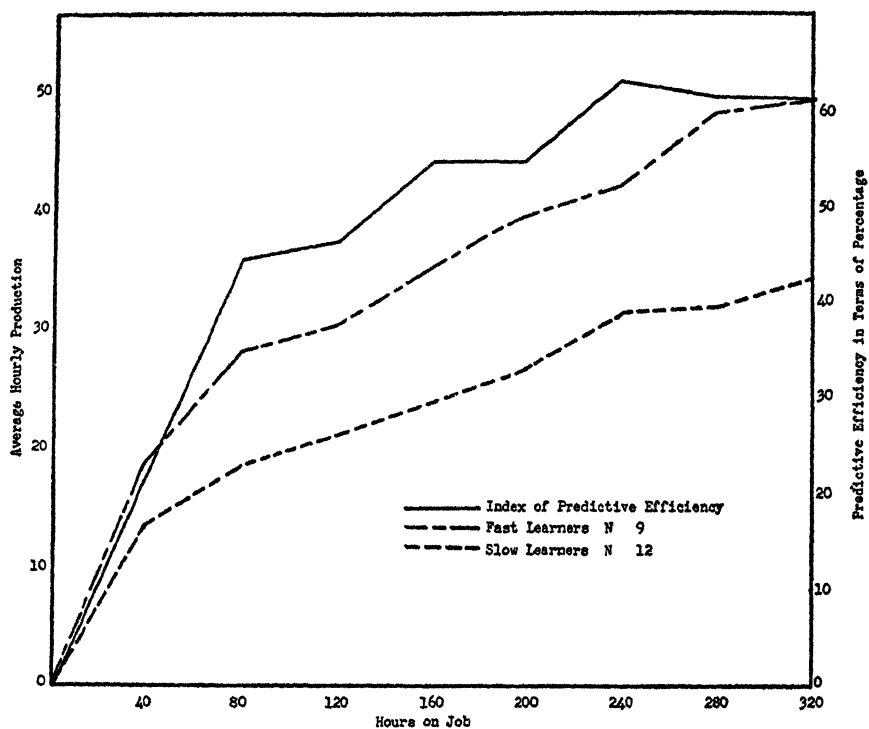



FIGURE 13 1 *Production of fast and slow operators and preductive efficiency*

TABLE 13 2

Means Production Standard Deviations, Difference between Means and Probability Values of Fast and Slow Groups for Each of First Eight Weeks on the Job

Group	*Period*							
	1	2	3	4	5	6	7	8
Fast								
Mean	18 4	27 8	29 7	34 9	38 7	41 6	47 3	48 9
S D	6 1	6 9	5 8	9 3	8 6	11 1	11 0	12 7
Slow								
Mean	13 3	18 2	21 0	23 1	26 5	31 0	31 6	34 0
S D	3 2	3 8	5 6	4 9	5 1	6 4	6 3	9 5
Difference, M_f—M	5 1	9 6	8 7	11 8	12 2	10 6	15 7	14 9
P Values (%)	<5>2	<1>0 1	<1>0 1	<1>0 1	<1<0 1	<5>0 1	<1>0 1	<2>1

ard production sooner but produce more than the slow learners at the end of each 40 hour period from the first throughout the eighth week The supervisor in this instance, if he decided to retain only operators who would make average production in less than average learning time, could have dismissed those who composed the slow group at the end of the fifth to sixth week of training He would still retain the majority of workers who learned the job on an average of about 2 months less time than the average operators in the total group of 21 employees

The differences between these groups are stable enough to use their production as standards against which to judge the performance of subsequent trainees similarly selected The amount of the difference between the two groups at the end of the first and the sixth periods could have occurred by chance only 5 times out of 100 The difference at the eighth period could have occurred by chance only 2 times out of 100, the differences at all other periods could occur by chance only 1 time out of 100 We would expect therefore, a new employee whose performance during the initial period of training was similar to that of the fast group to reach average production at an earlier date than would an employee whose early performance resembled that of the slow group

When to Decide

How early can the supervisor make his decision to retain or dismiss an employee in training on the basis of his performance in comparison with our fast and slow groups? In Figure 13 1 an additional line has been drawn to indicate the predictive efficiency of the performance of an employee at a given stage in training The supervisor by knowing how the employee performed during the first week of training could predict whether or not he would be fast or slow in reaching average production in a manner which would be more accurate by 20 per cent than would a chance prediction At the end of 2 weeks on the job, knowledge of an employee's production would increase the accuracy of the supervisor's prediction 44 per cent At the end of 6 weeks of training his accuracy would increase 63 per cent if he knew how the employee had produced during training The supervisor would not increase his accuracy of prediction from knowledge of how the worker produced during the remaining 2 weeks of the first 8 weeks on the job

This Index of Predictive Efficiency enables the supervisor to decide whether to retain or dismiss a worker in terms of the current labor situation If the labor market is tight, the worker could be retained for six weeks before deciding whether to release him However if the market is loose, the decision could be made at an earlier period A second factor which must be considered is the cost of the training If workers are scarce and training relatively inexpensive, one can keep trainees until failure is nearly certain, if the labor market is loose and training costly, it is uneconomical to continue training workers with lower probability of success

Technical Section

Careful statistical analysis is required in presenting any quantitative data When the number of subjects in a study are as few as those used in this study, generalizations are dangerous even after a careful statistical evaluation of data The industrial research man, nevertheless, must deal with relatively small numbers In this particular study the group of 21 included the entire number of operators trained during the period of record who remained on the job 15 months The author is well aware of the limitations as far as generalizations are concerned from so few cases, but believes the methods presented in the previous section and the statistical basis of this method to be presented in this section can be used profitably in industrial training [1]

The statistical significance of the difference between the fast and slow learners at the various periods in this study was tested by the method described by Lindquist (1,

[1] The writer appreciates the suggestion of Dr Robert J Wherry of Ohio State University which led to the use of the Wherry Doolittle Test Selection Method in this study

57) for testing the difference between means of small samples The confidence levels of these differences are shown in Table 13 2 As stated above, two of these differences are significant at the 5 per cent level, one at the 2 per cent level, and the remainder at the 1 per cent level

The Index of Predictive Efficiency (2, 459)[2] shown in Figure 13 1 is based on the shrunken multiple correlations between production for each of the first 8 weeks of training and the criterion of the total time required to reach average production The correlations were derived by a slight modification of the Wherry Doolittle test selection technique described by Stead and Shartle (2, 245–250) Zero order product moment correlations were computed between production during each one week training period and production in every other such period These correlations are shown in Table 13 3 All correlations except one (519) are significantly above zero at the 1 per cent level This correlation is significant at the 5 per cent level

The 21 operators were then divided as previously described into fast learners (N = 9) and slow learners (N = 12) Biserial correlations between the production at each period and the time required to reach average production were computed These are shown in the third from the last column of Table 13 3 The biserial correlations for the first, sixth, and eighth periods are significant at the 5 per cent level, the correlation for the third period is significant at the 2 per cent level The remaining correlations are significant at the 1 per cent level

TABLE 13 3

Zero Order Product Moment Correlations between Production of Each 40 hour Period with Every Other 40 hour Period, Biserial r s between Production Periods and Time Required to Reach Average Production, R s between Production Periods and Time Required to Reach Average Production, and the Index of Predictive Efficiency (E)

Production Periods	*Zero order Product Moment r's*								r_{bi}	R	E
	1	2	3	4	5	6	7	8	P *vs* T	ΣP *vs* T	
											%
1	1 000	793	655	730	806	781	770	519	599	599	20
2		1 000	793	741	848	793	754	565	836	834	44
3			1 000	888	781	812	800	655	760	842	46
4				1 000	848	818	916	787	792	887	54
5					1 000	860	921	640	833	894	54
6						1 000	866	591	629	929	63
7							1 000	780	842	924	61
8								1 000	679	923	61

The method developed by Wherry for test selection involving computation of a multiple shrunken correlation was followed, with two exceptions, in the computation of the correlations between the training periods and the time required to reach average production Wherry (4) has indicated how biserials can be used in multiple correlations The first exception to the method outlined by Wherry involves starting with the first 40 hour period rather

[2] It may be that the interpretation of the correlation coefficient suggested by Brogden (*Journal of Educational Psychology* February 1946, p 65) is more appropriate in this situation This author has established that the correlation coefficient "gives directly the ratio of the mean standard criterion score of a group selected by means of the predictor to that which would be obtained by selecting a group of similar size by means of the criterion itself This means that the average prediction of the best 9 men as retained at the end of 6 weeks would be 93 per cent of that attained by the best 9 men selected after 10 8 months

than with the period that has the greatest $\frac{V^2m}{Zm}$ ratio The second departure consisted in carrying out the computations for correlations for all eight 40 hour periods rather than stopping at the point where the shrunken multiple correlation begins to decrease The correlations which resulted from this technique include successively the effect of each training period while minimizing the errors involved in adding these practice periods The shrunken multiple correlations derived by this method are all significant at the 1 per cent level These correlations are shown in the next to the last column in Table 13 3 The Index of Predictive Efficiency (E) was computed for each of these correlations The E's are shown in the last column of Table 13 3

The E's show clearly that the predictive efficiency of knowledge of performance in a 40 hour period increased with the addition of each subsequent period up to the sixth (240 hours on the job) However, the increase in predictive efficiency is more pronounced from the first to the second period than between any subsequent periods The increase here is approximately 100 per cent The increase in accuracy of prediction when periods 1 through 6, inclusive, are compared with period 1, is over 200 per cent This means that the supervisor, when labor is plentiful and training costs low, can afford to release the worker who produces slowly during the first 2 weeks of training, however, when the labor is scarce or training costly the supervisor will want to retain these slower workers for a much longer period of training until he is more certain of final success or failure In any case, the knowledge gained from any time after the first 6 weeks of training will not increase his ability to predict speed of reaching standard production

The shrunken multiple correlations allow computation of weights to assign to the production of an individual for each of the 40 hour periods of initial employment It would be possible also through the use of weights to determine a production cutting score for each week of training The supervisor by comparing the weighted score of an individual employee with a cutting score for the period, can then make a decision to retain or dismiss the employee

The computations required for establishing cutting scores and indices of predictive efficiency are somewhat laborious Effective use of this method, moreover, would require revision of the cutting scores when any appreciable change occurred in working conditions, manufacturing processes, or quality of applicants accepted for the job Yet, once cutting scores are established and a basis for computing the level of achievement of individual employees is set up, it is easy to explain this procedure to supervisors and to use it in the work situation The increase in accuracy with which a supervisor could make his decision to retain or dismiss an employee through the use of the method described in this study would justify the labor involved in computing cutting scores and indices of predictive efficiency

References

1 Lindquist, E F , *Statistical Analysis in Educational Research* Boston Houghton Mifflin Company, 1940

2 Peatman, J G , *Descriptive and Sampling Statistics* New York Harper & Brothers, 1947

3 Stead, W R , Shartle, C L , et al , *Occupational Counseling Techniques* New York American Book Company, 1940

4 Wherry Robert J , 'Multiple Biserial and Multiple Point Biserial Correlation," *Psychometrika* 1947, Vol 12, 189–195

The Learning Curve for Flying an Airplane *

W N KELLOGG

The investigation was financed by the Civil Aeronautics Authority through the Committee on Selection and Training of Civilian Pilots of the National Research Council The data for this study were obtained in 1939 and 1940, but publication was necessarily withheld until after the termination of the war

The object of the present investigation was to examine the process of learning to fly It was directed specifically towards the plotting of learning curves and the study of the manner in which the student pilot eliminates his incorrect or erroneous responses as he masters the flying technique

PROCEDURE

Apparatus In order to obtain objective records, a special mechanism, known as the pilot response recorder was developed and installed in a Piper Cub Trainer This device is illustrated in Figure 14 1 The pilot response recorder weighs about 10 pounds and makes automatic graphic tracings of the extent and duration of the rudder, aileron, and elevator movements while the plane is in flight By means of a system of cams or wedges (*W*, Fig 14 1) the absolute extent of the airplane control movements is transmitted to the clockwork polygraph of the pilot response recorder in direct linear proportion The writing pointers are mounted on sleeves and move in a straight line across the paper Errors which are common in similar devices, such as the distortion introduced by the arcs of writing levers which are pivoted at a fulcrum, errors of changing air pressure within pneumatic systems, or the variation in the elasticity of tambours at different tensions, were eliminated by this method The entire apparatus was mounted in a concealed position behind the cockpit [1] It was therefore possible to keep the student pilot from knowing that records of his flying were being made at all

Sample records made by this device are reproduced in Figure 14 2 The lines show movements of the rudder, elevator, and ailerons which were traced during the process of making landings The first ground contact in each instance is indicated by the vertical broken line, so that, except for subsequent bumps, the portion of each tracing to the right of the broken line represents taxiing Time intervals shown on the bottom horizontal line are 10 seconds in length

Types of records made A standard course, which required about 10 minutes to fly, was laid out with fixed pilons on the ground The course included four left turns and three right turns Pilot response records for flying the course with records of the corresponding landings and take offs were made by *both student and instructor* after approximately every 30 minutes of flight instruction Periodic records were also taken of steep and shallow figure eights and of 360 degree glides to a landing

The weather control technique The object of having the instructor make flying records along with the student was to obtain some kind of a base or standard with which to compare the student's performance This base could not be a fixed one, but would be constantly changed or modified by variable weather conditions To cancel out this possible source of error, the instructor made the same maneuvers as the student, either immediately before or immediately after the student had made them Since the student's and the instructor's records were obtained but a few moments apart, over the same terrain, the difference between them could be regarded as a dif

* Reprinted from *Journal of Applied Psychology*, Vol 30, No 5, October 1946

[1] The pilot response recorder has been patented by Indiana University under the name of the airplane multiple control recorder

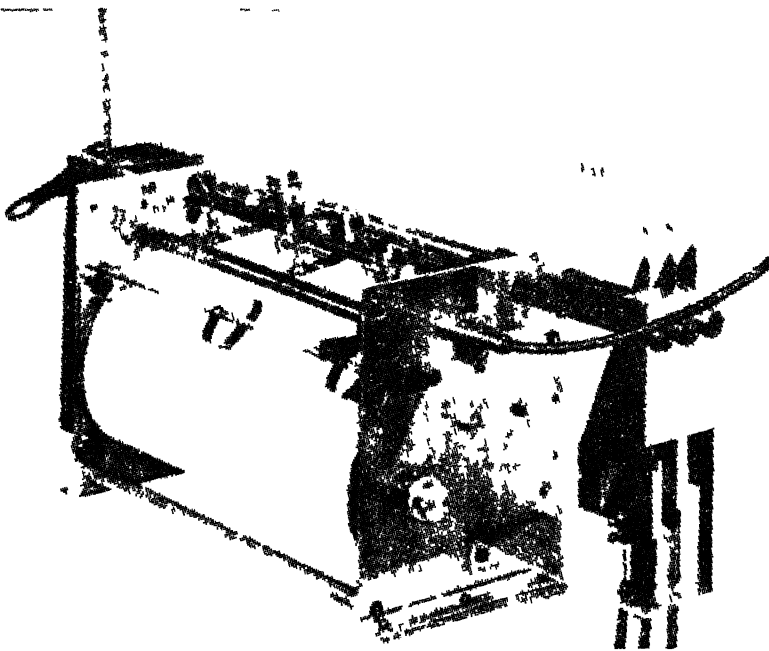



FIGURE 14 1 *The pilot response recorder is a light weight polygraph by means of which the movements of the airplane controls can be graphically traced A patented system of cams or wedges (W) transmits the control movements to the paper in linear proportion to their absolute extent Errors which might be introduced by the arcs of writing levers which are pivoted at a fulcrum by pneumatic systems or by the variable tensions of tambour diaphragms are eliminated by this construction*

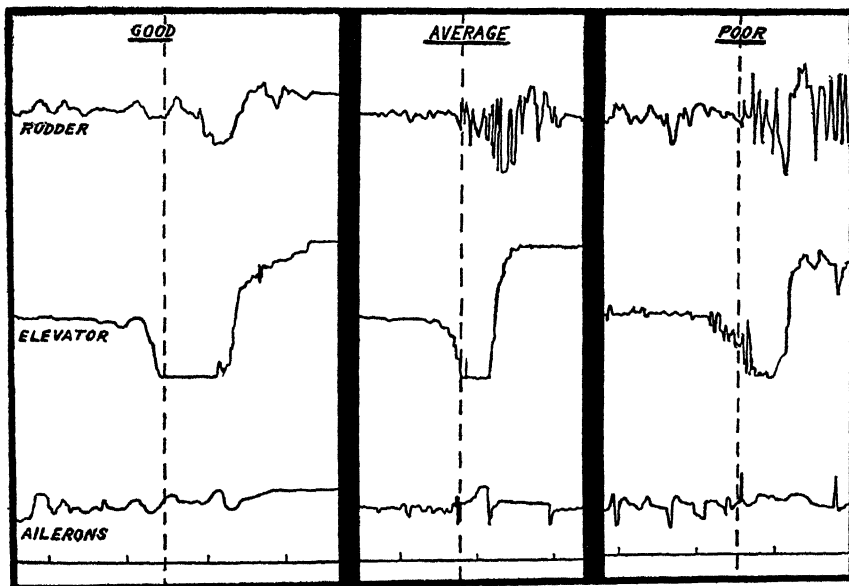



FIGURE 14 2 *The irregular lines show movements of the rudders, elevator and ailerons which were recorded during the process of making landings First ground contact in each instance is indicated by the vertical broken line The tracings to the right of the broken lines therefore represent taxiing Time intervals on the horizontal line at the bottom are 10 se[illegible] in length*

ference between the skill of the expert or finished pilot and the performance of the beginner

Every student record therefore, had paired with it the corresponding record made by the instructor under the same flying conditions To find what a student s errors were one compared the objective record of his flight with the appropriate control record made by the instructor This method has been called the weather con trol technique

Quantifying the data The graphic rec ords made by the pilot response recorder were measured by means of a special de vice known as a graphometer, which auto matically totals the vertical deflections or oscillations from the horizontal of any ir regular or wavy line [2] Readings from the graphometer converted to numerical form the total amount of movement of each of the airplane controls within any given time period By comparing the graphometer readings of the student and the instructor it was possible to tell at once which person moved any given control more or less than the other person, and *exactly how much more or less* he moved it

RESULTS

The results in this report cover the training of two student pilots Presented below

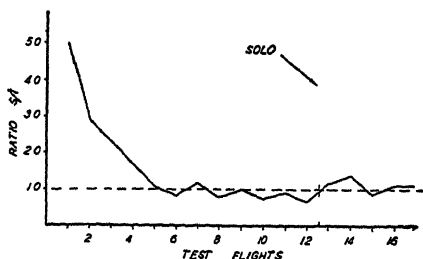



FIGURE 14 3 *Learning curve plotted from elevator movements of subject C showing the gradual elimination of over-controlling with practice in flying over a standard course Correct manipulation of the controls is represented by the horizontal line*

[2] W N Kellogg A Device for Measuring Kymographic Records, *Journal of Experi mental Psychology,* 1936, Vol 19, 383–385

are a few selected items which seem to offer the most promise for the analysis of the learning process

Course records In Figure 14 3 is shown the learning curve plotted from grapho meter readings of the elevator movements of student pilot C, during his flights over the standard ground course The points plotted are all ratios of the amount of elevator movement made by the student (S) divided by the amount of elevator movement made by the instructor (I) The curve includes 17 test flights or, roughly, 12 hours of instruction (17 half hour periods plus 17 ten minute periods of course flying) Student C made his first solo flight between test flight numbers 12 and 13

Since the points on the graph are all ratios, one can tell at once that student pilot C began his flying by moving the elevator about five times as much as the instructor moved it He was therefore over controlling very badly A ratio of 1 0 (indicated by the broken horizontal line) would mean that the student moved the elevator the same amount that the in structor did within the same time period It will be seen from Figure 14 3 that stu dent C gradually eliminated his elevator over corrections so that, after the eighth test flight, he was not far from the instruc tor s performance

In Figure 14 4 is shown a similar curve for student pilot C, but one which is a composite or combination of the move ments of all three of the airplane controls It appears from this learning curve that the student pilot on the whole moved the controls less than the instructor moved them This is indicated by the fact that the level of the curve is most of the time below the ratio of 1 0 Comparing the first part of the learning curve in Figure 14 4 with the first part of the curve in Figure 14 3 one may infer that since subject C over controlled so much with the elevator he must have undercorrected with the other controls As a matter of fact, this individual was much too limited in his rudder movements, as the graph of the course records for the rudder (not presented here) dem onstrated

Records of landings One of the most

difficult maneuvers which the new pilot has to perfect is the maneuver of landing It is, moreover, a maneuver from which many records can be easily obtained and one which must remain highly practiced with the pilot as long as he flies an air plane It should be clear also that in the maneuver of landing the elevator plays by far the most important part A good land ing is actually made only with the elevator and throttle (unless flaps are used) The rudder and ailerons should not be em ployed except in the approach to the field and in correcting for bumps in rough air

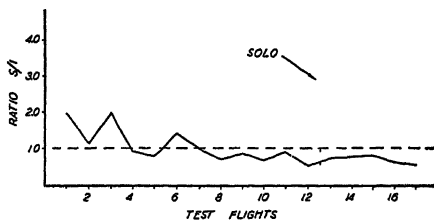



FIGURE 14 4 *Learning curve showing re duction in the manipulation of all three controls as compared to correct or ideal use of controls which is indicated by horizontal line*

The perfect landing is one in which the stick is gradually drawn backwards (the tail lowered) as the plane loses speed in its landing glide In the case of a three point landing the stick should be all the way back at the moment the tail and land ing wheels make contact with the ground (see Fig 14 2) Poor landings are those in which there is too much *forward* move ment of the stick The student "pumps the stick back and forth as he tries to 'find the ground' Improvement in landings should therefore be shown by the reduction in forward stick-movements with practice

In order to get at this problem, pilot-response elevator records were measured for a period of 15 seconds as the plane came into a landing A landing' was ar bitrarily defined by this means as the 15 seconds of flying time which ended with ground contact The learning curve plotted from such measurements, combined from the graphometer readings of the elevator movements of two subjects (C and P), is shown in Figure 14 5 Each point on the graph is a ratio of forward movements (F) divided by backward movements (B) of the stick—combined for two student pilots When the ratio is high (1 0 or 1 5) it means that the subject is pushing forward too much on the stick during his landings When the ratio is low it means that he is making few forward movements and that the landings are therefore good"

From an examination of Figure 14 5 it appears that there is a rapid improvement in landing skill for the first few hours of instruction, and that thereafter the prog-

FIGURE 14 5 *Showing improvement in the use of the elevator during landings only Composite learning curve for two subjects*

ress is slow—as in the mastery of any dif ficult skill

Conclusions

The following propositions seem to be justified by the limited data of this study

1 The objective analysis of airplane control movements can be used to show progress in the development of flying skill

2 By means of the pilot response re corder and the graphometer, the psychol ogist can tell which controls the pilot is manipulating incorrectly, in which direc tion his errors occur, and how great they are

3 The weather control technique seems to be adequate to cancel out variations in flying conditions

4 Learning curves for various maneu vers in flying are essentially the same as those obtained in the development of other skills

5 There is no evidence of plateaus in the curves of learning to fly, as plotted from the present data

PART TWO

Human Relations

The discipline of industrial psychology is concerned with both production and job satisfaction as a measure of ultimate efficiency It is extremely unlikely that true efficiency is possible without considering both factors The area of human relations in industry is becoming increasingly recognized as all important, and the psychologist's contribution is his recognition that assumptions, hypotheses, and theories are not accepted as if they were facts The experimental approach, although difficult, must be exploited before conclusions based upon facts in this area can be known

The three chapters in this section are not intended as complete coverage of the entire area, rather they are merely highlighted as examples of the wide variety of topics and problems Undoubtedly, the key to the area is to be found as a result of a better understanding of worker motivation The chapter on "Motivation, Related Factors, and Production" emphasizes the need for knowledge gained as a result of experimentation in this field Stating assumptions as if they were fact is likely to be misleading Speeches, exhortations, and imputing motives cannot be considered as satisfactory substitutes for a real knowledge of worker motivation

Labor management relations appear to be dedicated to the removal of conflict between these groups resulting from different motivations stemming from different group identification, interests, socio-economic level, and the host of other factors that make individuals different Psychologists have been slow to apply their methodologies in this important and critical area This chapter is intended as a spur in this respect

From among the very large number of attempts to change the various aspects of the work environment and thereby make it more pleasant, the single topic of music in industry was chosen At least, limited claims are based upon findings as a result of more or less adequate experimentation recognizing the need for control groups As a result, the uninitiated and unsuspecting are not subjected to as much nonsense and exaggeration in extolling the virtues of the specific panacea

Chapter IV

MOTIVATION, RELATED FACTORS, AND PRODUCTION

Although motivation is a vital key to industrial production, a comprehensive and systematic understanding of the relationship is not available The evidence from psychoanalysis and other sources reveals the complexity of motivation, and

yet the usual superficial nonexperimental treatment of motivation in industrial settings has retarded rather than advanced a real understanding of why people work

Studies comparing employers and employees indicate that the former attach greater significance to financial incentives It may well be that this group, projecting its motives to others, emphasizes the financial role by attributing such desires to workers

Of course people need money, but the amount they need is surely relative to their social, educational, and personal attainments and aspirations Whether a person will produce more in return for more money is truly a moot problem, it is not simple or obvious

Morale, job satisfaction, attitude, emotion, and other terms, are variously interrelated with the motivation of an individual This applies to all segments of life including the part related to a job Further, the job never exists in a vacuum Any vocational counselor or psychotherapist has ample evidence indicating that an individual's motives are usually many and complex, and, as long as he remains integrated, this applies to his job, home, and all other aspects of his life

Possibly the greatest difference between a psychologist and a layman is the psychologist's awareness of the complexity of motivation He knows that the basis of conflict is the stress and strain of different needs pushing at the same time He knows that the same situation can cause a conflict in one person and not in another Accordingly, advice is useless Amateur psychologists do not understand this and they freely give advice in the form of "I'll tell you what you should do—" or, "If I were you—" or, "Why don't you—" Such persons, despite the best of intentions, can do damage

The awareness of individual differences makes it impossible for a psychologist to render snap judgments of a general sort This makes people misunderstand psychologists, and they sometimes seek advice from those claiming omniscient powers

Objective data are needed The more meaningful analyses can only result from studies of broad scope Hypotheses and theories are useful but eventually the establishment of facts is necessary before worthy contributions applying to motivation on the industrial scene can be made Unfortunately, valuable work in this field has not taken place in proportion to the importance of the problem Much of it has been of the piddling variety, of the polyanna kind in which basic knowledge is sacrificed for "good feelings" Most work in this area has been inadequate because of superficial or faulty hypotheses accepted as fact

There have been a few notable exceptions to the previous statement, an outstanding example is the worthy experiment conducted at the Hawthorne Plant of the Western Electric Company No student of industrial relations (by whatever name) can afford to overlook the encyclopedic knowledge unfolded in this series of experiments

The decision concerning which phase to select was most difficult, and the section presented is not any more significant than others of the series It presents unusual insight into the intricate, informal social organization that employees construct to protect themselves Management, unaware of this structure, misunderstands employee motives and erroneously offers financial incentives to little avail

In an attempt to understand why people work, a study, employing a novel

approach to the basic problem of motivation, was conducted at General Motors This study, entitled "My Job Contest," is reported in a monograph published by *Personnel Psychology* On the surface this was simply a contest in which employees submitted letters on the topic "My Job and Why I Like It" There were more than 5,000 prizes awarded the winners, including a Cadillac, thirty-nine other cars, and an assortment of washers, frigidaires, bumper jacks, rear view mirrors, and so forth The employees responded Almost 50 per cent of the 297,401 eligible employees entered the contest The letters submitted varied in length from one sentence to twenty typed pages, about 700 were in languages other than English

Obviously this was more than a letter-writing contest It presented the oppor tunity to analyze thematically the relatively unstructured reflections of employees It was possible to translate these findings into management action in accordance with the expression of employees, and to indicate the relative value of management from plant to plant

Fragments of this monograph are selected to give the reader some notion of the breadth and scope of this study If possible, however, one should read the entire report

Marrow's report is most interesting since he is both an industrial psychologist and the owner of a rather large manufacturing business As a businessman he is expected to be concerned with running a profitable business and as a psychologist he has insight into the importance of experimentation His report not only indicates that experimentation is neither expensive nor profit-consuming, but also that experimentation can solve problems in a practical and valid manner

Many of the studies in this area have been of the survey variety in which the data accumulated have been confined to the limits prescribed by the measuring instrument The results, therefore, lack a depth of analysis that more spontaneous expressions or biographical accounts could reveal Nevertheless, results have value insofar as conclusions are based upon data, rather than upon "common sense" or unjustified and unsupported assumptions Jurgensen surveyed job applicants' attitudes toward ten factors using the rank order method The findings, not necessarily in agreement with the typical views of management and union leadership on the importance of various motives for working, demonstrate, nonetheless, the value of research

Kolstad heads an organization primarily concerned with conducting employee attitude surveys Excerpts from material released by this company are included to illustrate the extent of differences in employee attitudes on many different issues

Production, one of the most widely used criteria to determine industrial ef ficiency, is used to determine psychological test effectiveness It is used as a basis for wage rates Many gadgets, gimmicks, and environmental changes are considered effective in proportion to their ability to increase production Further, the entire system of time and motion studies has for its primary objective the increase of production Fatigue and monotony are also considered important because of their adverse effect on production It is obvious that production is regarded by many as the all important key

Actually, much more research is needed before production figures and work curves are taken out of the theoretical realm Unfortunately, not too much attention has been paid to this basic problem For example, the presence of restricted production is widespread, and yet incentive systems assume that their introduction causes restricted production to evaporate This is not true

Years ago a daily work curve was conceived as having a "warm up," reaching a "peak" at mid-morning, and then showing a "drop" until lunch The afternoon was supposed to resemble the morning period except that it didn't reach the morning peak and dropped much lower at the end of the day Such a curve indicated the presence of fatigue If the curve was very variable and ended with a spurt, then monotony was supposed to be present At least one large firm, vending industrial music, bases a large proportion of its research upon such assumptions These curves have not generally been found to exist The Hawthorne studies found, rather, that a flat constant work period was characteristic of the daily work process

There is a great need for basic analysis of the characteristics of the worker and his rate of production Rothe, in certain respects, has been doing some of this work He finds that work curves take on many different forms and do not assume any characteristic predictable pattern

The value of Rothe's work lies primarily in measuring production of workers on similar jobs His study of the output rate of butter wrappers is a good illustration of the light, manual, repetitive work so common in industry, and helps to explode the myth of the daily "fatigue" work curve In Rothe's study of machine operators, the major points to consider are the extent of individual differences of workers in similar jobs, and the variability in production from worker to worker in different periods

Rothe's study on "Output Rates among Chocolate Dippers" synthesizes his research in this area and allows for speculation concerning the relation between production, individual differences, and effectiveness of certain financial incentives

Study 4 Bank Wiring Observation Room *

MILTON L BLUM

This part of the Hawthorne Studies attempted to obtain more exact information about social groups within the company The preceding study had progressed from the proposed guided interview to a more intensive type of unguided interview and then to a series of interviews with one person In it, the emphasis was on obtaining information from large numbers of employees The last phase of the program pointed to the need to go back to a study of the Relay Assembly Test Room type, in which information of an intensive nature would yield data on the social groups in existence The reports of two of the interviewers will serve as a good introduction to the fourth of the Hawthorne Studies

They [the employees] firmly believe that they will not be satisfactorily remunerated for any additional work they produce over the bogey, or that if they do receive more money it could only be for a brief period, at the end of which the job would be re rated Because of the belief that rates may ultimately be lowered if output is too great, there seems to be a tacit agreement among the members of this group to limit their production to the bogey requirements on each operation Seldom do they exceed the

* Reprinted from *Industrial Psychology and Its Social Foundations* by Milton L Blum New York Harper and Brothers, 1949

bogey by a large margin Most of the work is turned out in the morning in order that they can take it easy during the latter part of the afternoon When questioned as to whether or not their earnings would be greater if they turned out more work, they claimed that the difference, if any, would be negligible because the percentages made by the other groups tend to pull theirs downward To this general scheme all their attitudes and behavior are related

The leader in this group is one of the two group chiefs, undoubtedly a very significant factor in giving the group a strong feeling of security This supervisor, A, was at one time on the bench in the same group which he now supervises, but he refused to allow the change to alter his relations with the men From observing the group one can hardly draw a line between supervisor and employees It is obvious that he is very popular with them, no one has any adverse criticism to make of him He is very close to the men, keeping them well informed at all times as to the group standing in the department, i e, relative percentages, rates output, etc When asked why they consider him a good supervisor, his men replied with such statements as He knows his stuff, He s fair and impartial," 'He'll go to hell for you and make sure you get plenty of work' In short, all their statements implied a firm conviction that this group chief would protect their interests By way of contrast, while A was on sick leave, another supervisor, B, took over the group Toward B the employees expressed strong antagonism B is an older man, further removed from the interests and sentiments of his subordinates He is not quite trusted by the men and commands very little respect As one employee sized him up, "When he bawls you out, he is more nervous than you are ' This group is only vaguely conscious of the other supervisors in the department, in fact, a confusion of the supervisory ranks is quite evident For example, C, a section chief, has held the same position for a number of years, but the men cannot figure out what position he holds in the department, who reports to him, or what his duties are

An attitude common to this group, but existing in varying degrees of intensity, may be characterized as a lack of ambition and initiative and a complacent desire to let well enough alone Most indifferent is their attitude toward advancement, referring of course, to promotion or higher grade work Whereas it is usual in any group to find several employees striving to improve their position, here only one or two seem to be interested The others merely say, 'All we are here for is the old pay check' Sometimes they speak of the department as the Old People s Home" because, quoting one man 'The fellows get in here and don t seem to want to get away Take a fellow like me I have been on this job ten years If I was transferred out, I would have to start all over again and I would have a pretty tough time

In their group life they are congenial and happy go lucky This is obvious not only during rest periods but also while they work Side play is frequent, and good natured bantering constantly flashes back and forth During rest periods everyone either plays cards or stands by as an interested spectator, and in these games rivalry is always keen but congenial Several of the newer men express the consensus of opinion by describing their associates as 'a swell bunch of guys' [1]

Another investigator reporting on a different group tells a different story

A says B gets mad because he (A) does too much work 'B sometimes tries to do as much as I do, and whenever he can t make it he gets mad and swears about it Then he will go over to some of the others and say that I am trying to kill the bogey The girl assemblers in the group tell A that he should not stand for the treatment he gets from the group chief They tell him he does the most work and gets the least recognition

A mistrusts D because D represented

[1] Reprinted by permission of the publishers from F J Rothlisberger and William J Dickson, *Management and the Worker—An Account of a Research Program Conducted by the Western Electric Company Hawthorne Works Chicago* Cambridge, Mass Harvard University Press, 1939

himself as a sort of a supervisor to A and took the easy jobs when A first came to work here He is not friendly with E be cause E does favors for everyone but him His friends are an old man, G, and the girls When B was called to the office be cause his production was too low, A told him it was his own fault B then said, 'What! Do you expect us to come down here and slave?

B is 36 years of age, a rather stocky, well built, athletic type Although he sup ports his father and mother, their depend ence apparently serves to accentuate his own manhood He says that the supervisors are all satisfactory He knows them well because he has worked here so long He takes a senior position in the group and gets along well with everyone but A His attitude toward A is indicated by the in cessant 'kidding to which he subjects him He attributes to himself all the best char acteristics of virile manhood and attributes to A feminine characteristics He says that A is an hermaphrodite He demonstrated in the first interview how A swings his hips and carries himself like a woman B thinks that A works hard because he is 'dumb," and that nobody likes A because he does so much He explains that A sits all by himself (in reality he sits next to B) and nobody will talk to him, so all he can do is work from the first whistle to the last B was once offered a position as supervisor which he refused [2]

The need for a more systematic inquiry resulted in the selection of 14 male oper ators who were to work under standard shop conditions These workers were ob served and interviewed over a period of 6½ months, the study was terminated when work ceased because of the depres- sion This group of men were reluctantly assigned to a separate room By now the researchers knew that such a change is often of importance, however, it enabled better control of the study The observer was stationed in the room, he was to as- sume the role of a disinterested spectator but was not to set himself off from the group He adhered strictly to the follow ing rules (1) Give no orders and in no way demonstrate authority (2) Do not take part in arguments (3) Do not enter into conversation nor seem overanxious to hear about what is going on (4) Never violate confidence of employees

The observer was asked to note the for mal organization of supervisor and em ployees, and also all informal groupings of the men Further he was to observe the interrelations of these two types of organ izations The interviewer did not enter the test room His function was to gain insight into the workers attitudes, thoughts, and feelings whereas the observer was to de scribe the actual verbal and overt behavior of the group Working thus together, these two were to gather data from this group concerning the department, the company, and the community

The workers in the Bank Wiring Ob servation Room study consisted of three groups 9 wiremen, 3 soldermen, and 2 inspectors Each did a specific task but necessarily collaborated with the others This department was chosen because it met such criteria as (1) the sameness of the task, (2) exactly determinable output, (3) shortness of task (1 minute required), (4) work pace determined by operator, (5) assurance of continued work, (6) the ease of removing the group as a unit from the department, (7) the experience of the operators These criteria were similar to those used in the Relay Assembly study, but from this point on there was a differ ence

The men were invited to cooperate in the study The first week they worked or appeared to be working all the time They were cautious toward the observer When they complained to him about poor light ing, he told them that he had no authority and suggested that they refer all complaints to their supervisor It was 3 weeks before the men started to relax and behave more as they did in their regular department It was learned that these men did not think that either the group chief or the section chief had much authority The foreman spent little time in the room so they were relatively free from authority

The system of payment was a compli- cated wage incentive plan that had been instituted to promote efficiency by encour

[2] *Ibid*

aging production, it was also believed to be a fair means of apportioning employee income It was soon found that this wage plan was not working The workers defined a day's work as the complete wiring of two units and either they stopped before quitting time, or they paced themselves to last out the day No uniform explanation or reason was forthcoming for this definition of a days work by the men, but it completely invalidated the incentive plan, as the following conversations prove

W_2 (After claiming that he turned out more work than anyone else in the group) They [his co workers] don t like to have me turn in so much, but I turn it in anyway

(In another interview) Right now I m turning out over 7000 a day, around 7040 The rest of the fellows kick be cause I do that They want me to come down They want me to come down to around 6600, but I don't see why I should If I did, the supervisors would come in and ask me what causes me to drop like that I've been turning out about that much for the last six months now and I see no reason why I should turn out less There's no reason why I should turn out more either

W_3 No one can turn out the bogey con sistently Well, occasionally some of them do Now since the layoff started there s been a few fellows down there who have been turning out around 7300 a day They've been working like hell I think it is foolishness to do it because I don t think it will do them any good, and it is likely to do the rest of us a lot of harm

Int Just how do you figure that?

W_3 Well, you see if they start turning out around 7300 a day over a period of weeks and if three of them do it, then they can lay one of the men off, because three men working at that speed can do as much as four men working at the present rate

Int And you think that is likely to happen?

W_3 Yes, I think it would At present we are only scheduled for 40 sets ahead In normal times we were scheduled for over 100 If they find that fewer men can do the work, they re going to lay off more of us When things pick up they will expect us to do as much as we are now That means they will raise the bogey on us You see how it works?

Int You say there is no incentive to turn out more work If all of you did more work, wouldn t you make more money?

W_4 No, we wouldn t They told us that down there one time You know, the supervisors came around and told us that very thing, that if we would turn out more work we would make more money, but we can t see it that way Probably what would happen is that our bogey would be raised, and then we would just be turning out more work for the same money I can't see that

W_5 There s another thing, you know the fellows give the fast workers the raspberry all the time Work hard, try to do your best, and they don t appreciate it at all They don t seem to figure that they are gaining any by it It s not only the wiremen, the soldermen don t like it either The fellows who loaf along are liked better than anybody else Some of them take pride in turning out as little work as they can and making the boss think they're turning out a whole lot They think it's smart I think a lot of them have the idea that if you work fast the rate will be cut That would mean that they would have to work faster for the same money I've never seen our rate cut yet, so I don t know whether it would happen or not I have heard it has happened in some cases though

W_6 (Talking about a relative of his who worked in the plant) She gets in here early and goes ahead and makes up a lot of parts so that when the rest of the girls start in she s already got a whole lot stacked up In that way she turns out a great deal of work She's money greedy That s what's the matter with her and they shouldn't allow that All she does is spoil the rate for the rest of the girls

Int How does she do that?

W_6 By turning out so much When they see her making so much money, they cut the rate

W_7 There's one little guy down there that

turns out over 7000 a day I think there s a couple of them And we have to put up with it [3]

The men devised various means of con trolling production Name calling and minor physical punishment were two of the more common ways of restricting out put Workers who produced too much were nicknamed 'Slave," 'Speed King, or 'Phar Lap' (a champion race horse of that year) They were also binged' A "bing is a very hard blow on the muscles of the upper arm The one who is hit never protests but is allowed to bing back

The men's concept of their average daily production was reflected in rather constant weekly production figures The men achieved this constancy by reporting more day work allowances than they were en titled to In addition they sometimes re ported more—or in some cases less—production than they had actually turned out The primary reason for this was to gain group acceptance

Three men always reported more work than they actually produced, and two re ported less, the others varied their reports A comparison of morning and afternoon production showed that the faster men slowed down in the afternoon whereas the slower men worked at a more even pace Briefly, the findings were that the men were restricting production in accordance with their definition of a working day, thus nullifying the validity of the wage incentive plan Interpersonal relations apparently were more important than the wage incentive

The group chief had certain difficulties In the first place he had to handle the day work claims of the men These claims were made to justify being paid at an hourly rate rather than on a production basis Company rules allowed such claims and they could be made for any number of reasons The group chief had either to accept these claims as justifiable or be arbi trary in rejecting them He chose to accept them and thereby gained the good will of his men It would have been difficult to prove any of the reasons given as being incorrect or unwarranted

[3] *Ibid*

Another problem was job trading The only excuse for this was physical inca pacity, as when a solderman developed a sore finger Determining how sore a sore" finger has to be made this a difficult claim to dispute Thus the group chief was sym pathetic to his men and steered a middle course, and he, in turn was popular with them During this study he was demoted because of business conditions and a group chief with greater seniority took over The new one placed great stress on conduct and efficiency The men thought that he was exercising more authority than was vested in him nor did they admit his au thority merely because he exercised it Certainly the first group chief with his leniency received more cooperation than the second one

The next representative of management was the section chief and since he super vised a number of groups he was never in close contact with any one group at all times His function was more managerial and he was considered to be more 'in the know' than the group chief Even though the men argued freely with him, they re garded him as having more authority than the group chief The assistant foreman, next in the management hierarchy, was listened to with respect but the men never argued with him If they disliked what he said, they waited until he left to voice their opinions The relations of the assist ant foreman to the group were pleasant The foreman was called the "old man" When he came in, conversation stopped and no one knowingly broke any of the rules The men showed apprehension while he was present

Considering the management employee situation, it was apparent that although communications traveled down in the form of orders, the two first line supervisors were likely to be questioned But there was a gap in communications on the way up from employee to foreman Conse quently the foreman and top management were unaware of the reasons for the failure of the financial incentive The fact that the men reacted differently than had been assumed made the incentive plan ineffec tual and was something the supervisory organization could not remedy

The relations between the employees were especially interesting The men worked according to their standard of production, but in addition they talked, argued, played games, matched coins or indulged in other forms of gambling, formed cliques, took sides, traded jobs, shared candy, insulted one another by belittling nationality and religion, and helped one another in their work They nicknamed each other Runt Shrimp,' Jumbo, and "Goofy" Their conversation ranged from work to women to horse racing In short, they did many things together, in addition to working

The connector wiremen, even though their rates might be the same as those of the selector wiremen, represented the elite Going on connectors was a step forward, whereas being put on selectors was regarded as a demotion The wiremen occupied a social position above the soldermen Job trading between them originated most often with a request from the wiremen The soldermen wore goggles which they resented, and the wiremen demonstrated their superiority by expressing disapproval when the soldermen did not wear them Lowest was the truckman who transported materials He was the butt of much horseplay

The inspectors belonged to a different group They were responsible to a different set of supervisors They were not an integral part of the group and were considered outsiders

A subtle manifestation of status appeared in the way the men dressed The foreman and assistant foreman wore coats and vests The section group chiefs wore vests but no coats The men wore neither coats nor vests When the men reported for interviews they did not put on their coats but the inspectors put on both coats and vests

During lulls in activity the men played games It was interesting to note that two groups always formed Group I consisted of four wiremen, a solderman and an inspector This group usually gambled Group II, not as completely set, consisted primarily of one solderman and three to five wiremen They preferred binging The third solderman and the other inspector were isolates that is, not in either group These groups or cliques carried over from games to job trading, quarrels over opening and closing windows, and friendships and antagonisms Furthermore, Group I regarded itself as the superior or front room clique They felt that their talks were on a higher plane, they ate chocolates rather than junk,' and they were less boisterous

There were four main determinants of clique membership (1) you should not turn out too much work (rate busting), (2) you should not turn out too little work (chiseler), (3) you should not tell a supervisor anything that would harm an associate (squealer), (4) you should not act officiously (this applied to inspectors and group chiefs as well as workers)

This intricate social organization served to protect the group both inside and outside Control inside was obtained through ridicule, sarcasm, and binging Protection outside was afforded by excessive daywork claims and constancy of production It has already been noted that management knew nothing about the group and its attitudes toward production and management rules until this phase of the study uncovered it All companies, large and small, have such a setup and under usual conditions they have no way of knowing about it

My Job Contest *

CHESTER E EVANS and LA VERNE N LASEAU

Analyzing the Letters to Build a Coding Structure

[Editor's Note This excerpt indicates how it is possible to analyze objectively data that are primarily subjective]

Since the basic research problems begin at this point, it would be well to state briefly the objectives of a research analysis and define the source materials with which the work must be done

The main objective of any analysis of the MJC entries would be to produce information that would be useful to the Division and to Corporation people in studies concerning employe relations It will be recalled that the third of the four objectives that General Motors had in mind when MJC was formulated was

(3) To collect material for the enlightenment and education of supervisory and management groups

The research objective itself was the last of these four objectives

(4) To obtain a body of data for the analysis of employe attitudes

The analysis was begun with the premise that it would be possible to study and analyze the human and personal documents to produce a significant reflection of employe thinking

A human and personal document, such as employes submitted in MJC, is a record of a person's thoughts when his mind is at liberty to discuss subject matter of interest or importance to himself

In MJC, the writer was given considerable latitude Except for the general subject, *My Job, and Why I Like It,* he was unconfined in what he chose to write about Obviously, the subject—and its application to the individual himself—provided a scope that could well include any segment of the writer's life experience, and thinking that he chose to describe

It is reasonable to assume that the typical entrant gave attention to what he considered important aspects of his experiences that might influence the judges to consider his entry favorably However, it is essential to remember that he retained a high degree of mental freedom regarding *what* he would write about

This is one of the most important aspects of MJC as a technique, for it produces a state of mind that is open and undirected The technique elicits for consideration all sorts of ideas, experiences, and theories on the part of the individual In this unconfined state, concepts of importance or interest tend to float to the surface Some concepts may be rejected for various reasons others may be used The important element is that this mental set facilitates the emergence of ideas or concepts that are psychologically meaningful to the individual The techniques of psychoanalysis and nondirective counseling depend for their success in establishing a climate where the patient can be free to express his unconscious desires and reveal his inner personality

Since MJC appeared to be conducive to a state of mental relaxation and since the entries themselves gave evidence of its existence, it was apparent that this raw material was a collection of human and personal documents This meant that an effective analysis of the MJC entries would provide highly reliable indications of the most important thoughts of General Motors employes regarding their jobs and related experiences that result from their association with General Motors

In comparison with other possible source material that could be used for exploring employe attitudes, MJC had some notable advantages and some troublesome disadvantages When related to the source material provided by various types of

* Reprinted from *Personnel Psychology Monographs* No 1, 1950

structured, formalized questionnaire and personal interview approaches, MJC appeared to have great advantages in the purity of reflection of the employe's attitudes and opinions However, the interpretation and analysis of so much unstructured material was quite difficult, and in a large quantity of such material, it added up to a formidable barrier from a research standpoint

On the other hand, the more formalized questionnaire or interview approach so limits and confines the respondent s answers that the violence done in thus restricting the free flow of his mental processes does not seem to be offset by the clear cut quantitative tabulations that result from adding up Yes or No answers

For some time prior to the beginning of MJC, thoughtful consideration was given to the basic problems of content analysis that were inevitable for quantifying the narrative data produced by MJC Conferences were held with experts in the fields of education, opinion and attitude research, social psychology, psychiatry, and political science, on how the MJC content analysis could be performed The most startling conclusion resulting from these various conferences was the fact that there was little precedent for the job ahead However, by accumulating the thinking of experts from all related fields, it was possible to map a tentative experimental design for the necessary content analysis work

Since nothing concrete could be accomplished until the actual entries were available, the pre contest planning could deal only in generalized terms Concrete steps were taken after the entries began to arrive Accordingly, as every 10th entry from the first 10,000 was pulled aside, 5 typewritten copies were made of it Even this procedure had to be discarded as the entries began to flood the office, with the result that it became necessary to photostat the last 400 of our original 10 per cent sample, so the entries would not be delayed for the initial routine records processing of the judging procedure

With a sample of 1,000 entries, careful studies were made to

1 Prepare the screening criteria necessary for the judging procedure Reading charts were prepared to assist in the elimination process

2 Construct a coding manual, based on a content analysis of the most frequently recurring themes discussed by the entrants The construction of the reading charts is discussed in detail in the Judging Procedure booklet referred to previously

In forming the coding structure, copies of the 1,000 sample were studied carefully by 5 independent groups Each group submitted a list of themes on which they felt the entries could be quantitatively analyzed By collating the various lists of suggested themes, it was possible to produce a list of over 150 prevailing themes, or coding categories Careful study of this list reduced the themes by about half, on the basis of their frequency of recurrence in the sample The list of themes finally selected consisted of 75 items, with provision also for coding the number of mentions of the Division" and of 'General Motors To check the number of entries that carried negative connotations, provision was made for a code— negative mention' To isolate quickly the entries making PS Comments, a coding category was provided that was called— backside mention' This made a total of 79 coding categories (thematic codes) against which the individual entries would be rated

[EDITOR S NOTE The following two tables present some of the results obtained from the content analysis of the letters Further, they indicate the types of satisfactions workers report The second table compares a specific company division with the entire company and shows that generalizations are subject to variation even in the same company]

The fact that the results on verifiable themes validly reflected the conditions as they actually existed gave evidence that the technique was sensitive to true differences between the Divisions as viewed by the employes Thus, it was possible to overcome the bias which initially had appeared inherent in MJC Although the employe could discuss only the *positive* factors about his job, his *lack of mention*

TABLE 16 1

How 69 THEMES Were Regrouped into 18 for Total General Motors

	Regrouped Themes	% *Mention*	*Original Themes*	% *Mention*
1	The *income* I get and the things it provides for my family and me	52 2	Wages and Salary	40 7
			Benefits Derived from Wages and Salary	24 0
2	The satisfaction and pleasure I get from doing an *interesting and important job*	50 8	Important Job	15 1
			Job Description	9 4
			Suitable Placement	5 4
			Attitude Toward Work	33 7
			Comparison with Other Jobs	5 6
3	The *pride* I get from being a part of such a *company*	49 4	Pride in the Company	32 2
			Pride in Product	25 4
			Pride in Building a Good Product	0 7
			Pride in Community Relations	3 4
			Pride in America	7 4
			Have Relatives in GM	1 6
4	The cooperation and team spirit of my *fellow workers*	48 9	Teamwork, Cooperation	18 7
			Fellow Workers	36 9
5	The ability and consideration of my *immediate boss*	47 9	Supervision (boss, foreman)	47 9
6	The fair treatment and policies of the *management*	47 5	Management	31 3
			Employe Employer Relations	4 6
			Fair Treatment	14 0
			Personnel Policies	6 0
			Personnel Department	3 6
			Non Discrimination—General	0 5
			Non Discrimination—Race, Nationality	1 3
			Non Discrimination—Religion or Creed	1 0
			Non Discrimination—Sex	0 2
			Non Discrimination – Age	0 6
			Non Discrimination Physical Handicap	0 8
			Veterans	8 4
7	The good tools, equipment, and *working conditions* provided me on this job	47 3	Working Hours	8 5
			Tools, Equipment and Methods	16 5
			Modern Plant or Office	3 4
			Air and Temperature	7 1
			Lighting	5 9
			Cleanliness	15 6
			Lockers and Showers	3 2
			Wash Rooms	5 7
			Comparison with Other Companies	7 9
			Good Working Conditions	1 4
			Cafeteria	10 1

TABLE 16 1—*Continued*

Regrouped Themes	% *Mention*	*Original Themes*	% *Mention*
7 —*Continued*		Parking Facilities	3 2
		Plant Location and Transporta tion	6 0
8 The feeling of *Security* I get from working for a stable company	35 6	Security	21 6
		Stability of Company	22 9
9 The *chance to get ahead* and the training and education provided to help me	35 2	Training, Education, Experience	21 7
		Opportunity for Advancement	21 7
10 The *benefit plans* provided, such as Hospitalization, In surance, and Bond Savings	34 2	Insurance Plans	28 5
		Hospitalization Plans	14 1
		Savings Plan	8 5
		Pension Plans	2 5
		Leaves of Absence	1 1
11 The emphasis on *safety* and the *medical facilities* avail able to me	32 7	Safety	22 5
		In Plant Medical Facilities	20 0
12 The paid *holidays* and *vaca tion* I get each year	16 6	Paid Holidays	3 0
		Vacation with Pay	16 1
13 The social, cultural, and *recreational facilities* pro vided for me	15 9	Recreation—General	4 4
		Sports	4 0
		Hobby Clubs	1 0
		Open Houses	0 3
		Christmas Parties and others	1 0
		Picnics	1 3
		Contests	5 4
		Employe Publications	2 6
14 *Personal achievement* grow ing out of my years of service with the company	14 2	Years of Service	6 9
		Personal History	6 6
		Success Theme	1 8
15 The *steady work* my job has provided	11 1	Steady Work	11 1
16 The opportunity which the *Suggestion Plan* gives me to capitalize on my ideas for improvement of methods, safety, etc	10 1	Suggestion Plan	10 1
17 The opportunity to enjoy the benefits of the *Free En terprise* system	7 9	Free Enterprise	7 9
18 Recogn t on [illegible]	[illegible]	[illegible]	[illegible]

TABLE 16 2

Percentage Distribution of 58 Themes in All Divisions and in Division # 48 [2]

Theme Name	*All Divisions Mention (% of Total Entries)*	*Division % Mention*	*% Difference from all Divisions*	*% Difference of the Difference*
Supervision	47 9	44 9	−3 0	−6 3
Associates	41 2	39 7	−1 5	−3 6
Wages	40 9	35 6	−5 3	−12 9
Work Type	33 7	35 3	1 6	4 7
Pride in Company	32 2	34 7	2 5	7 8
Management	31 3	27 7	−3 6	−11 5
Training Education, Experience	28 7	26 9	5 4	25 1
Opportunity for Advancement	25 6	25 8	4 4	20 6
Insurance	23 7	25 5	−3 2	−11 1
Security	22 8	24 3	2 8	13 0
Pride in Product	22 8	22 6	−3 0	−11 7
Pride in Stability of Company	21 5	21 8	−1 0	−4 4
Benefits from Wages	21 5	20 9	−2 8	−11 8
Teamwork	21 4	19 4	0 8	4 3
Pride in Important Job	20 2	17 9	2 8	18 5
Safety	18 6	15 8	−7 0	−30 7
Tools, Methods, Equipment	16 6	14 7	−1 9	−11 4
Steady Work	16 3	14 7	3 6	32 4
Fair Treatment	15 7	14 5	0 5	3 6
Paid Vacation	15 1	14 3	−2 0	−12 3
Non Discrimination	14 3	13 3	1 8	15 7
Recreation	14 0	11 6	2 7	30 3
Medical Facilities	11 5	11 2	−9 0	−44 6
Cleanliness	11 1	10 3	−5 4	−34 4
Suggestion Plan	10 5	10 3	0 2	2 0
Job Description	10 1	10 2	0 8	8 5
Comparison—Other Companies	9 4	9 9	2 0	25 3
Hospitalization Plan	8 9	9 6	−4 7	−32 9
Company and America	8 7	8 1	0 6	8 0
Working Hours	8 5	7 9	−0 6	−7 1
Free Enterprise	8 0	7 3	−0 7	−8 8
Savings Plan	7 9	6 8	−1 9	−21 8
Comparison—Other Jobs	7 5	6 5	1 1	20 4
Personal History	7 2	6 3	−0 3	−4 5
Pride in Years of Service	7 1	5 7	−1 4	−19 7
Parties and Picnics	6 6	5 7	3 9	21 7
Suitable Placement	6 2	5 6	0 2	3 7
Air and Temperature	6 0	5 6	−1 6	−22 2
Cafeteria	5 9	5 2	−5 3	−50 5
Employe Relations	5 8	5 0	0 4	8 7
Personnel Policies	5 4	4 7	−1 3	−21 7
Lighting	5 4	4 6	−1 6	−25 8
Pension Plans	4 6	4 1	1 6	64 0
Modern Plant	3 5	3 4	0 0	0 0
Plant Location & Transportation	3 5	3 1	−2 8	−47 5
Personnel Department	3 4	3 0	−0 5	−14 3

[2] The data in column 1 of Table 16 2 will not tally with the data for the same themes in the right hand column of Table 16 1, since Table 16 1 presents data on the cards unmatched to the vital statistics This matching out process resulted in the loss of 13,153 cases A further discrepancy results from the fact that Table 16 1 treats 69 themes, whereas Table 16 2 shows the recombination of all 77 themes into the final determined 58 themes

TABLE 16 2—*Continued*

Theme Name	*All Divisions Mention (% of Total Entries)*	*Division % Mention*	*% Difference from all Divisions*	*% Difference of the Difference*
Washrooms	3 2	2 6	−3 2	−55 2
Information Services	3 2	2 5	−0 1	−3 8
Success Theme	3 1	2 4	0 6	33 3
Pride in Community Relations	2 6	2 0	−1 5	−42 9
Paid Holidays	2 5	1 9	−1 2	−38 7
Parking Facilities	2 4	1 7	−1 5	−46 9
Leaves of Absence	1 8	1 7	0 6	54 5
Pride in Building Good Product	1 8	1 7	0 9	11 3
Seniority	1 3	1 6	−0 8	−33 3
Locker Rooms	1 1	0 9	−2 3	−71 9
Rest Periods	0 8	0 4	−0 9	−69 2
Open House	0 3	0 04	−0 3	−86 7

of certain factors was of significance in relation to his job and his Division If a particular activity or function was not being carried out, or if facilities were adequate and the local management was not communicating effectively with its employes, they did not select those themes as things to talk about Although the individual entry did not give evidence of the short comings of a particular Division as a good place to work, the cumulation of all the entries pointed rather clearly to things that were not talked about as much as would have been expected in terms of distributions across General Motors

Thus, MJC becomes a potent tool to uncover areas where Divisions are doing an outstanding job in employe relations and also to point up rather sharply those areas in which they are not performing in accordance with the other Divisions of General Motors

SUMMING UP

[EDITOR'S NOTE This section cogently summarizes the meaning of this study to management]

The tool that has been developed in the 'My Job and Why I Like It' contest is not a perfect tool nor is it easy to handle The material we obtained through its use will not give all the answers on what makes the American worker tick," but our observations have convinced us that the contest is the most helpful tool for the determination of employee attitudes *and* the development of effective employe relations policies we have found available so far

One basic problem of which all of us working in the field have been acutely aware is how to relate the study of employe attitudes to the facts of concrete conditions, policies, and practices in the plant We can find out the basic attitudes of the employe toward his job and his work, and we can find out, specifically, general categories of satisfaction such as security, pay, and recognition Unless, however, we can tie down these general categories to specific policies and their effectiveness, we cannot bridge the gap between research on employe attitudes and the formulation and administration of effective employe relations policies

All research work on employe attitudes has as its purpose the guidance of management thinking and practices toward a constructive policy, that is, toward a policy that will establish in the employe s mind a positive attitude toward his work

To accomplish this, we must know not only what the worker thinks about a particular practice, but also those policy areas in which positive action would have a marked effect on employe attitudes Some seemingly important aspects of a job may

have little effect on attitude while certain minor things may have high emotional leverage and considerable effect on employe attitudes What is it the worker considers relevant? What are the areas in which he is receptive to the right policy, either consciously or unconsciously, and what makes a policy right?

The worker's projection of his job into the MJC entries brought into clear and sharp focus the policies considered relevant by the worker Employe acceptance of MJC reflected a high level of sincerity and interest in this opportunity given them by management to express themselves There were frequent references to MJC as the first opportunity the employe had to tell management how he felt about his job

The supervisors also found MJC a useful and potent opportunity for additional contacts with their employes One supervisor told the story of an employe in his department who resisted making an MJC entry In following up this employe, who had over a period of time been somewhat uncooperative in other matters it developed that the employe's resistance to MJC and to the supervisor grew out of an unexpressed grievance he had nurtured for a long period of time—antedating the present supervisor In conversations with the supervisor, the employe was able to talk through the problem to his complete satisfaction Their relationship has since been more cordial It cannot be stated too strongly that MJC would have failed miserably without the cooperation and support of first line supervision The fact that they freely gave MJC their support gives additional evidence that in its essential nature MJC is an effective technique for improving employe relations

Obviously, any study of employe attitudes, whatever the technique, must be of such a nature that it does not upset the organization, otherwise it can only have harmful results, aside from being ineffectual But even the survey that is most conscious that it operates on a living body may cause disturbing reactions in the work force and particularly among first level supervision

On the one hand, management may be likely to take for granted the favorable things that result and to resent the unfavorable things as a criticism from some one not conversant with their operating problems MJC however, made sense to our practical operating people because it phrased in the worker's own terms an evaluation of management efforts in the field of employe relations The action patterns to which they have committed themselves are such that we believe the success of MJC as a total employe relations program is well established

It is not enough to say that General Motors has good employe relations The continual goal toward which we will strive is to have better and better employe relations

[Editor's Note Since the letters written by employees form the backbone of this study, it was considered desirable to select a few samples to indicate variation in theme, style, length, and so forth, and also to illustrate the human interest aspect that was packed into the study The letters are printed exactly as written]

Entry No 05 4159

My job has been nearly continuous for over 21 years at the Janesville Fisher Body plant That alone makes me appreciate my work but there are so many other *good* things that have materialized from my job that, it alone, seems small We own two homes that were acquired by Fisher and Oldsmobile earnings, have had 3 Chevrolet cars and Modern appliances in our home that were paid for by Fisher Body earnings in the 21 years on the job General Motors also made it possible for my older son to acquire a technical education at the G M Tech at Flint He has been an engineer for Fisher 10 years I could not have given him the education if G M *had not helped* My work at Fisher Body has been in the glass dept all the time I have been there With the exception of the war years I was a machine operator in the tool room for Oldsmobile during that time I am now working at glass salvage for Fisher and am also a utility man on the glass installation lines Its interesting work and I like it very much

My fellow workmen are a fine bunch of boys always ready with a good word and a smile I like to work with them The men who have been on supervision through the years have been helpful in a good many ways in my personal problems both in and out of the shop and I m proud to call them friends When the time comes to go home on my last day I will regret it very much

The services that are provided in the plant such as Group Ins Paid Vacations, Medical care in the plant, Cafeteria serv ice, Safety measures Labor saving appli ances, a clean healthy shop and personal advice given gladly, makes for a lot of sat isfaction In business dealings outside the plant the prestige of General Motors has given us better services from business and professional people and also a better standing with them personally It is a satisfac tion to me to do the best work I can because G M quality has always been at the top and we must keep it that way Some day buyers are going to be more particular and if we keep our products at top quality our sales department will have many repeats My job with Fisher Body has been pleasant and stable and has made a better and fuller life for my family and myself since I have worked there and now I want to say Thanks a lot for every thing'

Entry No 23 0700

Dear G M C

I'll write you a few lines two let you know why I like my job The Forman not all ways coming around saying thing there no need of I like the good wages I get Then I like all my fellow workmen I like name of working for G M C I like the hours I work I like the two 15 mi Rest Periods each shift I like the way we all do are work at G M C I like the job I am doing

Entry No 38 0795

Till the contest was announced I had more or less taken *My Job* for granted like heat & cold, sunshine & rain & summer & winter, but since I ve leaned back & looked at it from a personal angle I've formed some new ideas After a man s been on a job for better than 14 yrs as I have, it grows to be part of him like his wife & kids, his home, his lawn & shrubbery, auto mobile & furnace *My Job's* a lot like home & family, lots of pleasure, lots of grief, but either one or both will keep a man up on his toes year in & year out The comparison between home & job is fairly even all down the line Maybe it s a sick youngster at home & the doctors all out on calls & maybe at the shop it s a ship ment of deviation stock that taxes a man s wits & gives him a battle all day long to stay in the limits Then there are days when the work rolls out like hay & ma chines lay right on the money from whistle to whistle without any adjustment And at home the kids will bring home straight A s on their report cards & we'll have pop corn & apples & a fire in the fire place and maybe sing a few old songs (The kids like the new ones better) Oh *My Job* and my home are a lot alike Funny thing I get just as homesick for *My Job* as I do for my home if I go away for a while And I hate to lose time from *My Job* The other day I got up feeling a little bit under the weather so my wife asked me why I didn t 'sit this one out' But why should I? Nothing contagious & nothing wrong with me that would prevent me from get ting out a good day's work If I did feel worse on the job I could drop down to the hospital & our nurse would fix me up She s an old neighbor I've known since I was a little shaver Why I d trust her farther than I would most M D s & that s no knock on the medical profession either If I ever get seriously ill on *My Job* they ll have me at St John s in 15 minutes We feel we have a personal stake in St John's anyway GM donated half the funds for the new wing and the remainder was raised in the shop & town by popular sub scription With Blue Cross a man doesn't have the horror of hospitalization for him self or family that he had when the whole financial load was right up on his back At that I don t want any of it, especially the preventable variety *My Job* and the con stant preaching of safety has made me highly conscious of hazards around home, nails in boards, brooms on the basement

stairs and the like. Even the kids have a strong touch of it. "Mother put some merthiolate on this place" is a familiar cry around the homestead. I'm just as bad a safety crank as my boss. He's an old time machine man & remembers some of the horrible things machinery can do to a man that don't play it safe. Handy to have an old head around when I get into a crack that's a little too deep. Of course I'd rather whip my own trouble. Once in a while that turns out to be a good financial deal too. I've got several nice suggestion awards just for helping myself out of trouble. That's another thing about *My Job* that I like. By eliminating scrap, annoying, difficult & awkward operations I can not only improve my working conditions but get extra pay for doing it.

Overheard the neighborhood kids talking about jobs the other day. Mine said "My dad's a Delco-Remy man" & you know that gave me quite a boost. Didn't know I felt so strongly about the old place. I don't believe I want him to have *My Job* when I check out though. Hope he can take a shorter cut than I did. I'd like to see him in GM Tech some day. There doesn't seem to be any limit to where a sharp young fellow with technical training & an eye for the future can go with GM these days. Besides I don't feel like turning over *My Job* to anybody for quite a while. Reminds me: a bunch of us were down at the barber shop waiting our turn & the discussion got around to an acquaintance who had retired at 65. One of the boy said he hoped he didn't have to wait that long to be turned out to pasture. That's where I got in. I told them I just hoped to have enough left on the ball when I was 65 to keep on running *My Job,* and I'll bet that if I still feel that way when the time rolls around that *My Job* will be there for me, because I have a hunch that GM will still be operating at the same old stand.

Entry No. 40-0764

Why do I like my job? Each additional year that I am on the payroll at this plant more convincingly proves that I would find it rather difficult to find another firm able or willing to match the employment conditions I now enjoy.

To begin, the mention of years cannot but fail to sharpen the realization that I need but glance about me while on the job to see fellow workers whose years would deprive them of employment of their choice, ability, and experience at most places of employment, visible evidence to bolster my feeling of security. To me, security plus good wages are the primary requisites for a good job. From these two fundamentals stem all the numerous qualifications a job must have for it to be said that the job is a good job.

I like my job because it gives me a feeling of security. This feeling of security hinges on the fact that General Motors exemplifies security itself. This security enables them to find customers for their products, who might otherwise turn to another source of supply. Customers prefer to buy products backed by a reputable company. More satisfied customers means more profits. More profits enable a company to enlarge their scope of activities to benefit their employees. Most of these activities and benefits cost money, therefor I like my job because General Motors is a corporation that is making money and consequently able, willing, and does make these activities and benefits available to me.

Let me paint a word picture of a typical day at work. Specifically, my job is operating automatic screw machines, a semiskilled type of production work. Incidentally General Motors gave me my start on this work. My desire to work, at something I thought would be interesting and absorbing enough to enable me to ignore muscular fatigue, and for a higher rate of pay granted for labor requiring a little more than average skill and ability was recognized and I was launched on my new job that has now been my regular work for fourteen years, all for General Motors. The knowledge that this has actually occured enables me to enter the plant with a feeling of elation. Perhaps I may someday wish to change my job and armed with the above knowledge I feel confident I will be given the opportunity to embark on the new job if an opening exists, and retain it if my ability warrants it.

I enter the plant and while awaiting the sound of the bell to start work I can loiter in an immaculately clean dining room, to chat with fellow employees, engage in a friendly game of cards, indulge in refreshments dispensed for my pleasure, or simply sit around The bell rings and I proceed to my machines, walking along aisles free from an accumulation of dirt and debris, aisles flanked by machines periodically painted as required to give the plant an appearance of good housekeeping In my plant this good housekeeping is doubly insured by periodical inspections by a group of the higher executives that put down in writing any deviations of the good housekeeping rules and compare departments with departments and register the scores for efficiency in this respect I like this It scores another point why I like my job

Finally I get to my machines They are good machines, the last word in machines for that kind of work They are kept functionally perfect by a well staffed machine repair crew Should a mechanical problem arise it is immediately tackled by engineers of ability This means a lot to me I do not like to work with men groping for an answer for their problem Inefficiency is contagious and likewise innefficiency promotes a desire to be efficient yourself

I step up to the machine and press the starting button At once I am protected by emergency medical care in case of injury by facilities and staff members equal to a fine hospital, which in fact is maintained on a small scale on the premises I like to carry the assurance of this in my mind I like my job the better for it

My machine is in operation It is automatic I can look about me My fellow employees are busy Who are they? Why they are good, clean, law abiding American citizens You must be of this calibre to work in my plant I feel I am with friends, people of my kind, men you can depend and trust I have a feeling of well being No nervous tension This is conducive for efficient plant operation I like to work in an air of efficiency It stirs me to efficiency

I see my foreman He moves about the department performing his various tasks He knows what he is doing He has been trained for his job His ability gives me a sense of security Apple polishing foremen beget animosity and confusion The department loses its stability which in turn is reflected on the company s progress There goes your sense of security The company is going downhill But no need to fear that He has been chosen for the job because of his ability He understands the problems of his men He has a willing ear and an understanding heart

You have a suggestion for improving the job If it seems reasonably sound it is given a try It clicks You are rewarded for your effort It stimulates you to similar effort It stimulates your fellow workers to attempt to also reap your good fortune

You have certain rights Some things are forbidden Past experience has proven whether it is right to do this or wrong to do that Your foreman knows the rules and applies their interpretation alike on you and your fellow workers This is justice A feeling of justice is invaluable You like the management for their sagacity in seeing that justice prevails Injustice breeds contempt and hatred I hate to be stepped on and to be deprived of my just rights The right to justice is my legacy as an American citizen I hate to see my fellow man unjustly stepped on I like to work where the management assures me by deed that my just rights are protected

The bell rings for lunch hour I am able to eat my hot lunch which is served by the cafeteria service with hands cleaned with modern lavatory facilities, plenty of soap and towels simplify the chore I do not eat with dirty hands at home and I do not have to do so at the plant The management rightly recognizes this and has made arrangements so that it will not be necessary for me to do so

This is pay day I am handed my pay check On it are figures that represent definite assets to me It shows that I am paid as much or more money for my work as I can get anywhere in my community I see some deductions listed There is one for an insurance premium I know that no insurance company will give me that much protection for that amount of money General Motors has many employees and the saving on the premium is possible because of a group insurance plan They have seen

fit to inaugerate this plan that saves me money

I see another deduction This one is for United States savings bonds I am able to save systematically and regularily The management charges me nothing for this service which costs them money for additional office and clerical help My bond is delivered to me, with no effort or expense on my part I value this service I value this service the more when I realize that most concerns have discontinued that which was a wartime government request By continuing to deduct savings bonds at no cost to the employee General Motors is definitely proving that it has its worker s welfare at heart I like to work for some one that demonstrates this fact

This is the time I get two checks This one one is my vacation pay check Quite sizable too It represents a monetary token for sharing in General Motors profitable business

My days work has ended I proceed to cash my checks They are acceptable anywhere The checks have been issued by a reputable and a profitable concern Every one knows of it and the products it manufactures They are all good products In their field they stand at the top of the list of competitive merchandise No need to flinch when presenting a General Motors pay check The money is there to redeem it No need to flinch when A General Motors trade name is mentioned The company has been in business a long time The retailers of its products have been in business a long time The company with all its community business influences stands out prominently I am proud to present my check for cashing, anywhere

Now I am home I proceed to peruse the daily newspaper I daresay that hardly a daily paper goes to press that does not contain some reference to General Motors Corporation Perhaps it is an advertisement of their products It might be a financial statement of their dividends and earnings, and a healthy one at that That means that my job will be waiting for me tomorrow morning Not infrequently will you find an article announcing some invention or improved development by their wide awake research engineers Conspiciously on the page of community affairs you will always find the names of General Motors executives as active community workers

So my day is done One third of each of my working days is spent with General Motors I have been with them for fourteen years I expect to remain with them as long as I have to work for a living With More and better things for more people as their motto how can I go wrong You have asked me, Your job, And why do you like it You have my answer

Entry No 36 2000

When I was a boy I can remember my father coming home from the factory where he worked rather disgusted with his job Listening to him talk, I got the idea that a factory was a very poor way to make a living

When I hired in at Delco Remy I still looked at factories in that way but I needed a job and then too, I had noticed men that were working there and they seemed to be doing O K Factory conditions had either changed or else D R was different, because things Dad had talked about just didn t exist at D R At D R we had benefits dad had never heard of in a factory things that were really worth while, too

I know Group Insurance is worthwhile, its more than repaid me Vacation with pay isn't hard to take, either Those and many more have really sold me on D R I ve been there since "41 and probably will spend much of my future there I honestly believe the future will be bright for me

P S If the Cadillac isn t available in a darker color, just give it to someone else and I settle for a dark colored Buick

Human Factors in Production *

ALFRED J MARROW

Comparing the investigation of worker morale problems to medical science s earlier search for the cure for malaria, the author points out that progress remained at a standstill while therapy was focused only on the patient and his temperature But when such undreamed of measures as draining swamps (changing the environment) were tried, they proved to be the preventives that had been needed all the time This account of three group experiments into the causes of job dissatisfactions gives striking testimony to the value of the broader, environ mental approach to the problems of individual adjustment and interpersonal relationships in industry

The use of psychological techniques dur ing World War I gave great impetus, in the period that followed, to work in the field of applied psychology Unfortunately, the quantity of this activity frequently ex ceeded its quality aptitude testing almost assumed the proportions of a fad Inevi tably the over optimism and exaggerated claims of such enthusiasm led to disillu sionment, and the testing movement soon found itself in a position analogous to that of the public opinion polls after the presi dential election of 1948

As the 1920 s progressed with capable test researchers insisting upon more rigid standards, aptitude testing was restored to a position of scientific acceptability And, though industrial psychologists interested themselves in a number of other problems, their field continued to be largely a selec tion psychology

Within the present decade, however, there has developed a growing interest in the importance of interpersonal relation ships in industry Stemming largely from the field theory concepts of Kurt Lewin, many investigators are now approaching industrial problems from the social psycho logical viewpoint The individual is no longer to be studied as an isolated unit, but in relation to his environment, as he interacts with those about him The indus trial organization is considered a social unit, governed by laws of social interaction And if we would treat properly the per sonnel ills which beset us, we must first determine the social structure of our in dustrial organizations Through this ap proach lies the possibility of understanding and eliminating many of our personnel problems, and by so doing we create a healthy social unit

For the past decade, personnel research at the Harwood Manufacturing Corpora tion has been based upon these concepts Let us consider some of the theoretical considerations underlying our experiments

TIME PERSPECTIVE [1]

Morale is a much used and much abused term, defying adequate definition Never theless, we all know (too well!) what is meant by low employee morale

The level of morale is dependent to a large extent upon one s attitude toward the future The unemployed man who ex pects to find employment momentarily will maintain morale Once he loses hope com pletely, his morale drops We are all given, at times, to contemplation of the future We may be overly optimistic, unduly pessi mistic or alternate between these states, and rarely does the future unfold as we expect Nevertheless the psychological fu ture which an individual projects has a strong influence upon his moods and actions of the moment In like fashion are the psychological past and present of impor

* Reprinted from *Personnel* Vol 25 No 5 March 1949 American Management As sociation, New York

[1] The discussion in this section is derived from Kurt Lewin s chapter on 'Time Per spective and Morale, in *Civilian Morale* edited by Goodwin Watson New York Houghton Mifflin Company, 1942

tance, an individual's total time perspective will largely define his actions, emotions, and certainly his morale at any given instant

"Tenacity in the face of adversity" is a military definition of high morale, you must be able to 'take it' when the going gets rough And, while this may not be a generally acceptable definition, it contains an element which is applicable to the industrial situation

Persistence (a few degrees less determined than tenacity) is obviously a quality of value in an employee, and is definitely dependent upon an individual's time perspective Actually, there are two chief factors which make for persistence (1) the value of the goal, and (2) the outlook for the future If the psychological force toward the goal (the strength of the desire) is sufficiently great, and if there is a felt probability of reaching the goal, then we may expect persistent behavior

The felt probability of reaching the goal (time perspective of positive morale value) is a concept of some importance, as we shall see later To possess motivational value a goal must be attainable in the eyes of the goal seeker, otherwise it has no goal value, it is no goal at all Few of us continue to reach for the stars very long

Should a goal lack the probability of achievement, in terms of the individual's time perspective, the imposition of such a goal may very well lead to negative attitudes This is true for organized groups as well, and the negative attitudes developed may assume the form of organized aggression In this case, the group's time perspective serves to lower morale

Level of Aspiration and Time Perspective

In setting goals, or fixing upon a given level of aspiration, we are in effect expressing our wishes and hopes for the future The level of aspiration is therefore intimately associated with time perspective

A successful person will set his next goal somewhat higher than the last, but not so high that it cannot be attained In this manner, he is continually raising his level of aspiration He may have an extremely high ultimate goal, but he is realistic enough to approach it by achievable steps The unsuccessful person, on the other hand, will do one of two things (1) He will set a goal which is too low, an admission of defeat, or (2) he will set a goal which is too high, and then either make superficial gestures toward reaching it without really trying, or continue unreasonably to pursue the unattainable goal

Frequently a goal will not be accepted as realistic if there has been no demonstration that the goal is achievable We are all familiar with the "It can't be done!" response to a new piece rate Yet once the goal has been achieved, it acquires a social reality and is thereafter more readily accepted by others Their time perspective has been changed by the demonstration of achievement 'If he can do it, so can I'"

The foregoing theoretical considerations have been basic to the design of our research In demonstration, let us turn to the experiments

Turnover

Excessive turnover of personnel has always been a costly and disturbing problem, and so it became to us during the war In an attempt to understand the reasons for turnover, group interviews were held with supervisors The usual reasons were advanced moved out of town, took another job, no transportation, disliked the work, not enough pay, not enough overtime, etc As a follow up, exit interviews were conducted with all workers who left during a period of two months The workers mentioned a few additional reasons for quitting (e g, disliking their supervisors), but in general the reasons given were the same The opinions of supervisors and workers did not seem to give a final answer concerning the causes of turnover, nor were they of much practical help in prescribing remedies except for the elimination of rather specific grievances

The pressing problem of turnover demanded further psychological research From the viewpoint of factory management there were two purposes to the research (1) Why do employees quit their jobs? (2) What can be done to correct this problem?

Kurt Lewin, while visiting the Harwood Plant in 1944, expressed the belief that much of the turnover might be caused by a feeling of failure on the part of the em ployees who left the organization

To examine this possibility, the turnover of the previous month was analyzed The findings supported Lewin's hypothesis Of the 116 operators who were rating above standard production for the month, not one had left, but of the 211 who were rat ing below standard, 28 had quit during the month

The data also revealed the interesting fact that turnover increased as the worker approached the standard of an experienced operator (60 units per hour), that is, there was greater turnover among those who were approaching standard production than there was among those who were consid erably below standard This pointed to the possibility that the high quit rate of almost skilled' workers was caused by increasing frustration as they approached their goal It was theorized that this frus tration was caused by the conflict of two factors First, the strength of the worker s desire to reach standard increases as the goal comes within sight Second, the dif ficulty of improving production increases as the distance to the goal decreases, that is, the higher the level of production, the greater the difficulty of increasing produc tion Thus the conflict between (1) in creased desire to reach standard and (2) increased difficulty of reaching standard, seemed responsible for the frustration which led to a feeling of failure

The frustration failure hypothesis was explored further The employees who quit during 1944 were divided into seven groups, classified by amount of production at time of quitting For each classification the per cent turnover per month during 1944, based on the total number of employees in that classification, was computed

These data further substantiated the hy- pothesis It was found that the rate of turn- over increased as the learner approached the experienced level of 60 units per hour and decreased sharply once the success feeling of exceeding 60 units per hour was attained The monthly turnover at 30 units per hour (about half the minimum skilled level) was 1 per cent, at 45 units per hour it rose to 5 per cent, and at 55 units per hour (almost equal to a skilled level) it rose to 8 per cent On an annual basis, the turnover figures were equivalent to 12 per cent, 60 per cent, and 96 per cent, re spectively, for the three indicated numbers of units per hour Once the standard was achieved, the turnover rate dropped to 13 per cent

If the frustration failure hypothesis of turnover was valid as the data seemed to indicate, then we had the answer to our first question Employees quit their jobs (at least many of them) because of a sense of failure We were then faced with the second problem What could be done to correct the situation?

To eliminate failure experiences, insofar as possible, the training program was re designed The trainers were instructed not to put the slow learner on the defensive, but rather to accept without criticism the explanation which the trainee offered Trainees were encouraged to accept the factory situation realistically, with all its unavoidable difficulties and minor irrita tions In addition, though the trainee was told of the ultimate goal she was expected to reach, she was also encouraged to set weekly goals which seemed reasonable In this connection, the trainers were carefully instructed to recognize unrealistic goal setting, it was necessary to avoid over opti mistic levels of aspiration which might result in failure Trainees, in conference with the trainer, often set as many as eight or nine substitute goals during the training period Probably many more were set un- officially on a daily or two day basis (per- haps hourly) once the habit was formed

The results seemed to substantiate the hypothesis As a result of the trainer s en- couraging, sympathetic attitude, and the establishment of realistic substitute goals (as contrasted with the less realistic goal of standard production) the turnover rate dropped about 50 per cent for the entire plant More significant was the drop in turnover at the 'almost skilled' level, the high frustration period, turnover at this level dropped to 75 per cent in 1947 In 1944 it had been 300 per cent

EFFECT OF TRANSFERS

American industry has characteristically changed models and products as frequently as competitive conditions and engineering progress dictated When this happens, it is necessary to transfer workers to different jobs In addition, the high rate of turnover and absenteeism in recent years resulted in unbalanced production lines which also necessitated shifting of workers

The resistance of workers to such transfers was one of the major problems of the production staff When transfers occurred, the resistance would become evident in grievances concerning piece rates on the new jobs, high turnover, low efficiency, restricted output, and overt aggression toward management But there was no way of avoiding the transfers, plant operation required the changes in methods and jobs

Here again we seemed faced with a failure situation As the first step in the investigation, an analysis of turnover was made, comparing transferred operators with those who had not been transferred recently Data were gathered for the period from September, 1946, to September, 1947 Each group was divided into seven classifications according to production rating at time of quitting For each classification, the per cent turnover per month, based on total number of employees in that classification, was obtained

When the curves were plotted, it was discovered that both the level of turnover and the form of the curves were quite different for the two groups The non-transfers showed an average monthly turnover of about 4½ per cent, among the transfers, the average monthly turnover was approximately 12 per cent In comparing the curves, both groups showed a sharp drop in turnover for those beyond the 60 unit standard, this was consistent with the findings of previous studies that turnover was low among those who had experienced success However, whereas both groups had high turnover at the 'almost skilled level, again consistent with previous studies, the transfers also had a peak turnover at the lowest production rating, evidence that many of the transfers quit just after their production fell upon being transferred

How were these findings to be interpreted? The high turnover at the almost skilled level already had an explanation in the frustration failure hypothesis previously described But whereas the learners experienced greatest frustration at the level just below standard, the transfers experienced even greater frustration immediately after transfer It seems likely that the frustration resulted from loss of face, the contrast between the previous high status and present reduced status

Still another revealing finding was made in the investigation of resistance to transfer For the simplest type job in the plant the average learning time for beginners was five weeks Yet experienced operators, when transferred to this same job, required an average of eight weeks to reach standard Nor could it be argued that the skill habits on the old job were interfering with relearning on the new job Transferred operators rarely complained that they wanted to do it the old way, and time and motion studies showed very few false moves after the first week of change Thus the slow relearning of transfers seemed primarily a motivational problem In support of this was evidence obtained from interviews with transferred workers The group was characterized by a general pattern of low morale There were expressions of resentment and aggression against management but largely, the picture was one of resignation, with evidences of frustration, loss of hope of ever regaining the former level of production and status, feelings of failure, and a very low level of aspiration

All the standard approaches were used in attempting to solve the problem of resistance to change Special monetary allowances were made, the cooperation of the union was enlisted, layoffs were attempted on the basis of inefficiency very little was accomplished

Here again, special research procedures were required Starting with a series of observations about the behavior of transferred groups, a general over all program was devised The first step consisted of developing a theory to account for the resistance to change Next, a real life action

experiment was designed to be carried out as an integral part of plant operation Finally, the results of the experiment were interpreted in light of the theory, accounting for the experimental differences and providing effective means for overcoming resistance to change

The experiment consisted of selecting four groups of operators, three experimental and one control, selection was based upon similarity in efficiency and other important variables A change, comparable in degree, was then made in the operation of each group Thus the four groups became experimental transfers

The manner in which the operations of the groups were changed varied this was the experimental variable For the control group the usual factory routine was followed A meeting was called at which it was explained that the change was necessary because of competitive conditions, that a new piece rate had been set, the rate was explained, questions were answered, and the meeting was dismissed

The operations of the experimental groups were changed differently Included in the method of change was a forceful technique for making the group aware of the need for change, and an opportunity to participate in planning the details of the change Group 1 was represented in the planning by selected members, Groups 2 and 3 all participated in the planning

The results of the experiment were fairly clear The control group dropped in production immediately upon change, and by the end of the experiment showed no appreciable amount of recovery Resistance developed almost immediately There were marked instances of aggression against management, deliberate restriction of production, lack of cooperation with the supervisor Nine per cent quit during the first 15 days after the change Grievances were filed about the piece rate which, upon checking, was found to be even a little "loose"

The recoveries for Groups 2 and 3 were dramatic Both groups recovered to their pre change level of production the second day after change, and by the end of the experiment they had actually surpassed their pre change level by about 14 per cent They worked cooperatively with their supervisors there was no indication of aggression, and there were no quits during the 15 day period

Group 1 required more time to recover (possibly because of an unavoidable operational problem), but reached the pre change level by the 14th day after change, and by the end of the experiment had exceeded its pre change level Here, too, no quits were recorded One act of aggression was observed which was neither prolonged nor serious

In analyzing the negative attitudes which make for resistance to change, there seem to be four component forces in operation, these correspond to the goals of pay, security, status, and success When an operator is at a production level above standard, the situation represents high pay, security, status, and success—all of a positive nature, when at a level below standard, we have low pay, insecurity, lack of status, failure Because management's standard of production is well accepted, any action which tends to place the operator at a level below standard is strongly resisted

The success of the experiment seemed to be attributable largely to the fact that experimental transfers were given the opportunity to participate in planning the change, in planning their own work future Thus, where such external motivating forces as monetary rewards, management pressure and other means had failed, group involvement and decision developed internalized motivation for the accomplishment of a goal mutually desirable to management and worker

Group Standards as a Restraining Force

It has been demonstrated that a group member will tend to set his goals in accordance with the standards of the group An interesting illustration of the restraining forces which group standards bring to bear upon goal setting is found in the following report

In 1937, management was confronted with the problem of training hundreds of workers for a newly erected plant in the south It was assumed at that time that

skill was the major, and motivation the minor, determinant of the rate of learning During the first year of the plant's operation, a traditional training program was introduced The average output after a week s training ranged from 10 to 20 per cent of the standard production for skilled workers The trainees were informed at that time that their first week s rate of production, frequently attained at great physical and emotional effort, was only a fraction of what they would be expected to produce at the end of 12 weeks The disparity between their accomplishment and the stated future goal, an imposed goal, was so great that many workers expressed scepticism of ever being able to attain that goal Since the plant was newly organized, there were no skilled workers who were actually doing the job at standard speed, the goal therefore seemed impossible to attain The wages these learners received were already more than they had been accustomed to as domestics, farm hands, or waitresses, there was nothing either inside or outside the plant to give the higher standards social reality for the group As a result, despite the dissatisfaction of management, the learners were well satisfied with their progress Consequently the learning rate was slow, plateaus were common, and at the end of the 12 week training period the majority of the trainees were only producing about 50 per cent of the minimum standard for experienced workers Pressure methods were introduced to increase production Rewards and punishments of many kinds were tried The result was a frightening increase in voluntary quits Finally, after some 40 weeks, the first few workers reached the minimum level for rating as skilled

It had become clear by this time that an individual will slacken his efforts and set his goals far below those he could reach if group standards are low Conversely, he will raise his goal if the group standards are raised In other words, both the ideals and the action of an individual depend upon the group to which he belongs and upon the goals and expectations of that group

A comparison of the training period of the first year of plant operation with that of the second year illustrates the effect of group goals on the individual At the beginning of the second year a number of the older workers had already exceeded the skilled level of production Moreover, there had been a migration to the plant of a group of experienced workers from a nearby community Thus the group standards had been appreciably raised In addition, the changed reality of the goal had an effect upon learning and the level of aspiration As a result, the average training time of 34 weeks in 1937 was reduced to 14 weeks in 1938

So much for the vivid testimony to the great restraining force of a seemingly unattainable goal, as exemplified by the experience in 1937 and its sequel in 1938 Let us turn our attention now to the present

During the war a popular model was discontinued After a lapse of four years, this item was reintroduced At the time of the reintroduction, there were no longer any workers in the plant who had formerly made the product Consequently, it represented a new item for which there were no accepted group standards, just as it had in 1937 Since there were actually no workers doing the job at a standard speed, the goal once again, in 1948 as in 1937, seemed too difficult and unattainable As a result, the learning curve flattened out to the level which had prevailed some 10 years earlier when the plant was first established

Note that the average training time, which had been 34 weeks in 1937 and had been reduced to 14 weeks in 1938, has been still further reduced in 1948 to 7 weeks Yet on this one item, for which there are no group standards, there is every evidence at the moment of writing that approximately the same 34 weeks will be required to reach standard And so we have completed a full circle

There are several elements in common between the 1937 and 1948 situations In both cases, the job was new to the plant personnel In both cases, the goal seemed unattainable And in both cases, the group established standards which restricted production, though for different reasons In 1937 the group standard remained low be

cause additional pay did not constitute a motivating factor The 1948 low group standard is an intentional restriction of production, an aggression against manage ment

There have been many instances in the past when a new job required extended learning time because the goal seemed un attainable to the original group of learners And after the first group had attained the goal and given it social reality, succeeding groups were able to reach standard pro duction within a shorter learning period And so we predict that, once the present slow learners reach standard, succeeding learners on the same job will reach stand ard in much less time

CONCLUSION

These experiences emphasize that in in vestigating the general problem of job sat isfaction and morale, we must explore beyond the limits of the worker in his job We must do as the medical scientist did in his search for a cure for malaria So long as the therapy was focused on the patient s temperature, little progress was made But when such undreamed of measures as draining swamps and stocking ponds with fish were tried, they turned out to be the preventatives which had really been needed all the time Similarly, our investigations of the problems of human adjustment must be broader than the symptom and the pa tient We need to seek out the environ mental causes of conflict and dissatisfaction, its media of communication, the unhealthy social conditions which make for low re sistance to its transmission or persistence, and in general, its social pathology and therapy

What Job Applicants Look For in a Company *

CLIFFORD E JURGENSEN

All of us have heard persons say, Com pany X is the best company to work for in the whole state' Similarly, we have heard persons say I wouldn t work for company Y if it were the last place on earth ' Applicants frequently state in their interview for employment, I would like a job here because everyone says it s a good place to work These opinions which people have toward any company are im portant to that company, and are partic- ularly important in the case of a public utility They do much to establish and maintain good or poor public relations, and they make it easy or hard to build up an adequate pool of job applicants from which to select satisfactory employees

Favorable opinions of the company as a place to work are also important so far as our present employees are concerned It is a source of pride and job satisfaction to work in a company that is well thought of Happy and satisfied employees result in increased work output, improved quality, and decreased costs Job satisfaction also has an important bearing on labor relations

What are the factors by which persons decide whether a job is a good job or a company is a ' good company? Discus- sion with executives, supervisors, union officials, employees, and job applicants will bring to light a number of such factors This discussion will be limited to the fol lowing ten

Advancement (Opportunity for promo- tion)
Benefits (Vacation, sick pay, insurance, etc)
Company (Employment by company you are proud to work for)
Co workers (Fellow workers who are pleasant, agreeable, and good work- ing companions)
Hours (Good starting and quitting time

* Reprinted from *Personnel Psychology*, Vol 1, No 4, winter 1948

good number of hours per day or week, day or night work, etc)
Pay (Large income during year)
Security (Steady work, no lay offs, sureness of being able to keep your job)
Supervisor (A good boss who is considerate and fair)
Type of work (Work which is interesting and well liked by you)
Working conditions (Comfortable and clean, absence of noise, heat cold, odors, etc)

All persons will probably agree that these ten factors are important Opinions will differ however, as to the *relative* importance of the factors Within management it is not uncommon, for example, to hear two statements The most important thing to any employee is the size of his pay check, and security is more important to any employee than anything else Obviously, both of these opinions cannot be right Usually arguments on this subject are ended with no change of opinion although sometimes they are won by the man who is the best talker or by the man who has the position of highest authority

Similar disagreements arise in contract negotiations, grievance hearings, and other meetings between management and union officials Solution of industrial conflict is frequently stalemated because representatives are unable to agree on what is desired by employees Each representative, of course, is sure that he knows what employees want and that opposing opinions are wrong

It is interesting to speculate on the reasons for this diversity of opinion Two things appear to be particularly important in determining our opinion of what others want

1 We tend to attribute to others those interests and desires which we ourselves have

2 We tend to overemphasize those factors which we hear most about

Since V J Day, the Minneapolis Gas Company has been collecting information on what *job applicants* want most This has been done by means of a questionnaire containing the 10 factors and definitions previously listed Each applicant was asked to decide which of the 10 factors was most important to him, and to place a 1 in front of that factor Then he was to decide which was second in importance to him and to place a 2 in front of it He continued in this way until all 10 of the items had been ranked in order of importance to him Each applicant was told that there were no right or wrong answers and that he was to answer according to what *he* thought rather than what he believed others might think Applicants were not asked to sign their names although for research purposes they were requested to give their sex, marital status number of dependents age, salary, amount of education and main occupation

Nearly 4000 applicants for jobs were asked to rank these factors in importance to them The results showed Job Security to be most important Opportunity for Advancement second, and Type of Work third Fourth in importance to these job seekers was Pride in the Company This finding indicates a need for a greater effort to sell the company as a place to work and as an employer the worker can be proud of Instead of ranking first or second, as might have been expected Pay tied with Co workers for fifth place among the ten factors These findings characterized the group as a whole There are in addition important differences in the relative significance of these factors between types of applicant (men or women, older or younger, sales, clerical skilled semi skilled or unskilled) These differences among groups repay further study

This report is based on data collected in the 2 years immediately following V J Day in August, 1945 Three thousand seven hundred and twenty three applicants are included 3345 men and 378 women This report will be limited, in the main, to a discussion of job preferences of men Not only are more data available than for women, but the typical company employs far more men than women

Findings for Men Applicants

Security was the most important factor so far as men are concerned As might be expected, it was more important to mar

ried men than to single men Its importance decreased as extent of education increased Mechanical workers were more interested in security than were clerical workers, and sales applicants were relatively least interested in security The emphasis given security appears well warranted

Advancement was second in importance Applicants wanted opportunity for promotion Whether they wanted promotion on the basis of merit or seniority cannot be determined from these data

The importance of advancement became greater as extent of education increased Sales, clerical, and skilled applicants were most interested in advancement, and they were followed by those who are semi skilled Unskilled laborers were less interested in advancement than were any other groups It is interesting to note that those who were already at the top of the occupational ladder were most interested in further advancement, and those at the bottom were least interested in advancement It would be interesting to know whether ambition to progress results in a high job level or whether increase in job level results in an increase of ambition Some light is shed on this from the responses of high school students who considered advancement less important than did most occupational groups This appears to indicate that ambition to progress may often follow rather than precede actual advancement

Type of work was, surprisingly listed third in importance Single men were more interested in type of work than were married men, and the importance of this factor decreased as number of dependents and age increased Type of work increased in importance as the job level increased, the order from low to high being unskilled, semi skilled skilled, clerical, and sales It is interesting to note that the persons who worked on the least pleasant jobs were those who were least interested in type of work

Company was listed fourth in importance It was relatively more important to applicants with a large number of dependents, older applicants, and those on the higher job levels Much more importance was assigned to working for a good company than would be expected on the basis of the small amount of time and effort expended by most companies to sell the company to the employees These data indicate the need for greater activity of this type House organs, bulletin boards, employee induction manuals and other communication lines can be used to give information to employees and increase their pride in their company

Pay was listed in the fifth position There was considerable discrepancy between the importance pay is usually believed to have and the position actually assigned it by applicants Surprisingly pay decreased in importance as the number of dependents increased and as age increased It became more important as the extent of education increased Sales applicants rated pay higher than did applicants for other positions The viewpoint toward pay can be summarized by saying that relative to other factors it was considered by applicants to be only average in importance, and that it was least important to those applicants who might be assumed to be in greater need of high pay

The discrepancy between typical opinions regarding the importance of pay and its importance as actually given by almost four thousand applicants needs some explanation

Perhaps pay is more important than admitted by job applicants in this and other similar studies, a pay increase is usually the first demand of any union On the other hand, let us look at the experience of companies which have tried to forestall unionization by giving large and numerous wage increases History indicates the futility of these attempts, and in many cases unionization occurred quickest in those companies paying the highest wage rates Furthermore, the most intense labor strife frequently has been in those industries and companies which have paid the highest wages Pay obviously is not a panacea which will solve all controversies

From the employee viewpoint there are several reasons which may explain the undue emphasis on pay All employees would like to secure more pay than they have secured in the past, and if they believe they can get more they may try to do so

TABLE 18 1

Average (Mean) Ranks Assigned Job Factors by Various Sub Groups of Men
(1 0 indicates maximum importance 10 0 indicates minimum importance)

		Marital Status			Dependents						Age								Education					
	Total	Single	Married	Other	None	One	Two	Three	Four	Five or more	Under 20	20 24	25 29	30 34	35 39	40-44	45-49	50 or over	8th grade or less	Some H S	H S diploma	Diploma plus	College attended	College degree
Advancement (2)	3 6	3 7	3 5	3 0	3 7	3 6	3 4	3 3	3 7	3 2	4 2	3 6	3 4	3 3	3 5	3 7	3 3	4 2	4 3	3 7	3 5	3 3	3 3	3 2
Benefits (10)	7 4	7 5	7 2	7 5	7 5	7 3	7 3	7 1	7 1	7 1	7 5	7 5	7 4	7 2	7 2	7 1	7 6	6 9	6 6	7 1	7 5	7 5	7 8	8 1
Company (4)	5 0	5 6	4 5	4 4	5 6	4 7	4 6	4 3	4 0	4 2	5 9	5 6	4 9	4 4	3 9	3 8	3 4	3 6	4 5	4 7	5 3	4 6	5 3	4 6
Co-workers (5 5)	6 0	5 7	6 2	6 0	5 8	6 2	6 2	6 2	5 7	5 7	5 6	5 9	6 2	6 4	5 8	5 9	6 0	5 5	5 5	5 6	6 2	6 4	6 2	5 8
Hours (8)	6 9	6 6	7 2	7 2	6 6	6 8	7 3	7 3	7 5	7 0	6 1	6 6	7 1	7 2	7 6	7 4	7 4	7 2	6 5	6 8	6 8	7 3	7 1	7 7
Pay (5 5)	6 0	5 8	6 2	6 4	5 8	6 1	6 1	6 2	6 6	6 2	5 3	5 6	6 0	6 3	6 9	6 8	6 5	6 9	7 1	6 5	5 8	5 7	5 1	5 3
Security (1)	3 3	3 7	2 9	3 2	3 7	3 0	2 8	2 9	2 7	3 1	4 1	3 4	3 1	2 9	3 0	3 1	3 2	3 5	2 9	3 0	3 1	3 1	4 1	4 6
Supervisor (7)	6 1	6 2	6 0	6 6	6 2	6 1	6 2	6 1	5 4	5 5	6 1	6 3	6 3	6 0	5 7	5 5	5 5	5 5	5 3	5 9	6 3	6 5	6 4	6 0
Type of work (3)	3 7	3 5	3 9	4 1	3 4	3 8	3 7	4 0	4 7	5 3	3 8	3 5	3 4	3 8	4 0	4 8	4 7	4 6	5 0	4 2	3 5	3 2	3 0	2 5
Working conditions (9)	7 1	6 8	7 4	7 1	6 8	7 3	7 4	7 5	7 7	7 7	6 7	7 0	7 2	7 4	7 5	7 1	7 3	7 0	7 3	7 4	7 0	7 4	6 7	7 2
Number of cases	3345	1522	1665	98	1394	838	541	285	124	73	272	1171	842	452	253	128	77	87	314	908	1166	283	526	78

TABLE 18 2

Average (Mean) Ranks Assigned Job Factors by Various Occupational Sub Groups of Men

(1 0 indicates maximum importance 10 0 indicates minimum importance)

	Student	*Sales*	*Clerical*	*Mechanical* Total	Skilled	Semi skilled	Un skilled
Advancement	3 9	3 2	3 2 (2 5)	3 6	3 2	3 6	4 0
Benefits	7 7	7 9	7 8 (10)	7 2	7 3	7 2	7 0
Company	6 3	4 1	4 7 (4)	4 9	4 9	4 8	5 1
Co-workers	5 8	6 5	6 3 (6)	6 0	6 2	6 0	5 5
Hours	6 1	7 9	7 2 (9)	6 8	7 2	6 9	6 4
Pav	4 6	4 9	6 0 (5)	6 3	6 4	6 3	6 2
Security	4 6	4 2	3 2 (2 5)	2 9	2 9	2 9	2 9
Supervisor	6 4	6 3	6 4 (7)	6 1	6 2	6 0	6 0
Type of work	3 3	2 6	3 1 (1)	4 0	3 4	3 9	4 8
Working conditions	6 4	7 3	7 1 (8)	7 2	7 3	7 2	7 3
Number of cases	322	222	259	2252	287	1626	338

TABLE 18 3

Average (Mean) Ranks Assigned Job Factors by Various Sub groups of women

(1 0 indicates maximum importance 10 0 indicates minimum importance)

	Total	*Age* Under 20	20 24	25 and over	*Education* H S Diploma	College	*Occupation* Student	Clerical
Advancement (3)	4 8	4 8	4 7	4 7	4 8	4 4	5 0	4 8 (3)
Benefits (10)	8 2	8 3	8 2	8 0	8 2	8 5	8 2	8 2 (10)
Company (5)	5 4	5 8	5 2	4 9	5 5	5 5	6 6	5 0 (4)
Co workers (5)	5 4	5 9	4 8	5 3	5 7	5 1	6 2	5 3 (6)
Hours (8)	6 1	5 5	6 5	6 7	5 9	6 3	5 1	6 3 (8)
Pay (9)	6 4	5 8	6 7	7 3	6 2	6 3	5 3	6 8 (9)
Security (2)	4 6	4 8	4 7	4 3	4 4	5 3	4 5	4 5 (2)
Supervisor (5)	5 4	5 9	5 1	4 9	5 4	5 6	6 4	5 1 (5)
Type of work (1)	2 8	2 9	2 7	2 7	3 0	2 2	2 4	3 0 (1)
Working conditions (7)	5 8	5 3	6 4	6 2	5 8	5 8	5 3	5 9 (7)
Number of cases	378	183	120	75	216	93	95	213

Further, demands for greater pay are often used as a substitute for other wants which may be either conscious or unconscious Sometimes these other wants are those which the employee has learned through bitter experience are not seriously consid ered or acted upon if presented to man agement For example, an employee is not apt to present a grievance against a super visor because of the manner in which the supervisor says "Good morning!" to his subordinates Nevertheless, an accumula tion of minor factors of this type may cause employees to be depressed, anxious, and tense Since grievances of this type are not generally settled satisfactorily if presented directly, they are often presented indirectly in such form as a demand for higher wages

The demand for higher pay is given fur ther emphasis when employees are on strike Not only do the foregoing reasons

apply, but additional reasons enter into the picture No group of employees strikes against an employer unless there is intense feeling on both sides, and there is frequently a desire to hit the opponent in a sensitive spot The pocketbook is such a sensitive spot, and so employees frequently emphasize a pay increase even though the basic reasons for the strike are far removed from the question of wages

Striking employees may also overemphasize pay increases in order to arouse public sympathy in their favor Sympathy is particularly easy to arouse if the public can be made to believe that the workers involved are grossly underpaid Use of this technique has often resulted in the public holding serious misconceptions regarding wages paid in particular industries For example, few persons realize that miners are relatively high paid, both from the viewpoint of hourly pay and annual earnings In the year 1944 (latest figures available) employees of all private industries averaged $2,189 income, whereas anthracite miners averaged $2,494 and bituminous miners averaged $2,534 These figures are quite different than many persons would guess

In summary it appears that the relative importance of pay has often been overemphasized This is not to say that pay is unimportant or that substandard wages will not result in employee discontent It does mean, however, that other factors are of equal or greater importance, and they should be emphasized more in the future than they have been in the past

Co-workers were tied in importance with pay In general, the various groups were quite consistent in their rating of co-workers, though it was rated somewhat more important by unmarried men, and those with less than high school education Applicants for unskilled work tended to rate co-workers somewhat higher than did other groups

Supervisor was rated in seventh position, being almost tied with pay and co-workers The importance attached to the supervisor increased as age and number of dependents increased, and decreased as extent of education rose The importance attached by applicants to a good boss who is considerate and fair needs to be emphasized, particularly to the supervisors themselves Few persons realize that the supervisor is within a hair's breadth of being just as important to applicants as their rate of pay, and the implications of this fact have most assuredly been lost by the wayside

Hours were rated eighth in importance by job applicants Hours became relatively less important as education, dependents and age increased, and was less important for married men than for single men Hours also decreased in importance as job status rose All groups, however, rated hours as being relatively unimportant Considering the enormous improvement in hours of work which has been obtained, it would appear that the point of diminishing returns has been reached, and that more time might profitably be devoted to factors currently considered more important by the average job applicant

Working conditions were rated in the next to last position Although all groups considered working conditions to be quite unimportant, it was least so for those applicants who had the most dependents As with the factor of hours, far too much attention, relatively, has been given to working conditions Although this factor may have warranted the attention given it one or two decades ago, conditions have improved to the extent that workers are not particularly interested in further improvement This, of course, must be interpreted from the viewpoint of the average job applicant toward the average job

Benefits were relegated by job applicants to the last position Married men were slightly more interested in benefits than were single men, and the importance of the factor rose as age and number of dependents increased Benefits became less important as extent of education and job level increased In view of these results, it would appear worthwhile to review reasons for emphasizing the importance of benefits There certainly is no justification for emphasizing benefits on the basis of these applicant's desires, although emphasis may be warranted by employee need

In summary it can be said that male job applicants consider security, type of work and advancement to be most important

An intermediate position is given company pay, co workers, and supervisor Least importance of all is attached to hours, working conditions and benefits

Findings for Women Applicants

Women applicants differed from men Women considered type of work, security, and advancement to be most important (Those are the same factors as emphasized by men, but the order differs) An intermediate position was given by women to supervisor, company, co workers, working conditions, hours, and pay Benefits were rated as being least important

Attention should be called to the fact that although pay received a mean rank of 6 4 among women, it was ninth in order of importance

In general it can be said that women were less interested than men in advancement, benefits, company, pay and security On the other hand, women were more interested in co workers, hours, supervisor, type of work and working conditions These differences form a definite pattern Women were particularly interested in short range or temporary factors which increased the pleasantness of work, whereas men were more interested in the factors of greatest importance for a lifetime of work to support themselves and their families

Conclusions

These data have important implications in selecting, training and supervising employees For example, excellent results have been obtained by writing help wanted advertisements which emphasize those factors ranked highest by job applicants The job preference blank has been an exceedingly valuable tool when used in employment interviews

These findings may be even more valuable in determining personnel policies and conducting union negotiations Data indicate that many persons have erred considerably in their opinions of applicants job preferences, and it must be remembered here that future employees come from today s applicants Union agreements typically emphasize wages, hours and working conditions which factors did not turn out to be the most important in the opinion of almost 4000 job applicants In negotiation of union contracts, as well as in day to day relationships between management and union, considerable emphasis is placed on security and benefits Security would appear to warrant such emphasis, but benefits may not

Discrepancies between union demands and preferences of applicants can be explained by any or all of the following hypotheses (1) union officials do not know what employees desire most, (2) union officials are more interested in what they believe employees *should* have rather than what they actually want or (3) union officials are most interested in items which will sell the union to employees

Management has also erred in many respects The type of work being done by

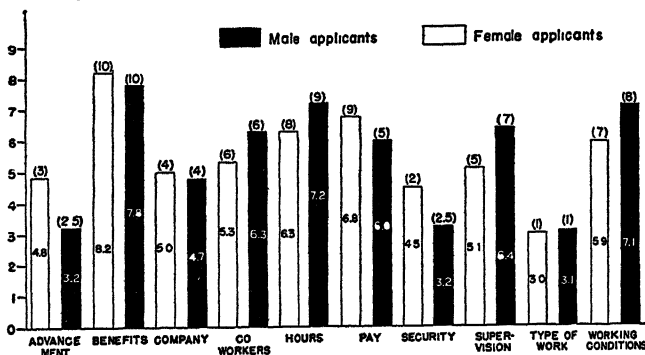



FIGURE 18 1 *Mean rank assigned to each of 10 factors by 259 male and 213 female clerical job applicants Numbers at the top of each bar represent the rank of that factor*

the employee is frequently considered by management to be of little importance to employees except in terms of gross classification such as sales, mechanical, clerical, and administrative These findings indicate the importance of managements making transfers or promotions only after discussion with the employees involved Although management may contend that few employees verbally object to undiscussed changes, this cannot be interpreted as absence of objections It is highly probable, in light of these data, that employees do not object to such changes because of fear of consequences or feelings of futility Such fears and feelings do not build favorable employee morale It would appear highly profitable for representatives of management and unions cooperatively to develop methods and procedures which would insure better placement of employees in the type of work which they would most enjoy

The high importance given by these applicants to working for a good supervisor warrants emphasis, particularly for those companies which consider the supervisor to be a necessary evil or a glorified worker It appears likely that companies could earn a large profit in terms of dollars as well as improved employee morale if they were to develop better techniques for selecting and training supervisors

Space does not permit discussion of many other implications of these findings Any interested person can find many important implications by studying the original charts upon which this report is based

In summary, too much emphasis has often been given factors which according to this study are considered relatively unimportant by applicants There would seem to be an excellent opportunity to devise principles and procedures which would result in greater job satisfaction on the part of employees, and consequently in improved quality of work, increased quantity, and lower costs

For the benefit of those who wish more details than could be given in this report, three tables are included which contain the original data on which this report is based

Excerpts From Employee Attitude Surveys (I) *

ARTHUR KOLSTAD

Departmental Variations in Attitudes

We stated in Bulletin No 10 that we have usually found just as large a range or spread in favorable percentages' for the departments within a company, as among companies

From a selected group of organizations we reported the extremes—the highest and the lowest percentages observed in employees 'favorable' responses to a number of questions For example, 97 per cent of the employees in Company R, but only 39 per cent of the employees in Company P rated their company as better than average as a place to work When the proportion of employees having a favorable opinion of the company approaches 100 per cent, it is evident that in most departments the favorably-inclined proportion must be large Actually the range for "rating the company above average' by departments in Company R was 78 per cent to 100 per cent

In Company P, however, the range was from 15 per cent to 90 per cent All employees worked under the same company policies, and all but a few warehouse employees worked under the same roof Yet in contrast to the majority of departments, there were 5 departments in which 80 per cent or more rated the company above average A good job had been done in

* Reprinted from *Employee Attitudes* No 11, Arthur Kolstad, New York

selling the company' to the employees in these departments But what about the job done in the 6 departments where less than one fourth rated the company above average as a place to work?

These two companies are among the extremes in our experience It may be more useful to examine the results in an 'average" company (Average" in that the overall results are average as compared with those in other companies for which we have made surveys) Seventy per cent of all its employees rated this company as better than average as a place to work Yet this company, too, has its strong points and weak points

The tables that follow will give an idea of the spread of favorable or satisfactory percentages among the departments of this "average' company Notice that the first items are ones which reflect attitudes toward the company and management If the administration of company policies is uniform throughout the company, less variation should be expected on these items than on those which reflect attitudes influenced directly by the supervisor's actions

A Department May Have Strong and Weak Points

A department may be strong in one area and weak in another Or it may be weak in only some one specific phase of a general area

A group of guards and watchmen expressed very favorable attitudes toward the company and its management, were very well satisfied as to pay, security on the job, advancement, benefit plans, general working conditions, etc In the field of supervision, relatively favorable responses were registered on all but one point—a large proportion objected to favoritism being shown in the assignment of tours of duty At the time of the survey, resentment had not spread beyond that toward the men who made the assignments

The employees in a manufacturing department expressed very unfavorable attitudes on all items relating to company management, but they rated their foreman very high, gave "above average responses to questions about supervision, and registered a high degree of satisfaction regarding pay and the work load

A clerical department, separated from the main office due to insufficient space, was, as a group, well satisfied with supervision in its various phases, with pay, promotion, general company policies, etc But the employees expressed the feeling that they 'did not belong —for which they blamed top management and were doubtful of management s interest in their welfare

Numerous illustrations could be cited of dissatisfactions not brought to management s attention through ordinary contacts and channels—dissatisfactions that could be easily remedied if discovered in time, but which could grow in importance if permitted to smolder

(*II*) *

Attitudes of Union and Non union Employees

We have often been asked whether there are any differences in the attitudes expressed by union members and those expressed by non union employees In answering this question, comparisons between unionized and non unionized companies are not valid Nor can we depend on comparisons between plants operating under union contracts and non unionized plants in the same multiple unit company, for we have found large differences in the attitudes expressed by employees at different plant locations of the same corporation, whether unionized or not And results based on a sample drawn from a number of different companies may be subject to error The ideal comparison would be one based on two groups of employees, doing the same type of work, under the same working conditions, under the *same quality* of supervision, etc , except that one group is working under a union contract and the other group is not

---

* Reprinted from *Employee Attitudes,* No 14, Arthur Kolstad, New York

Item	*The Two Lowest Departments* [1]		*The Two Highest Departments* [1]	
How many employees feel that their company is better than average as a place to work?	30%	44%	93%	100%
How many say that they are made to feel that they are *really* a part of the organization to a large extent (more than just to a fair degree)?	10	17	67	86
How many rate management's interest in the welfare of employees as being above average?	10	15	64	86
How many feel that top management is practically always fair with employees?	5	11	57	58
How many feel almost certain (or very sure) of holding their job as long as they do good work?	33	35	85	92
How many believe the company has a pretty good idea as to which employees are best qualified for better jobs?	33	40	82	85
How many know at least fairly well whether their boss is satisfied with their work or not?	17	30	82	100
How many are rarely, or never, "bawled out" or criticized when they do not deserve it?	38	45	95	100
How many usually (or always) get recognition or praise when they do some unusually good work?	0	0	58	67
How many feel there is no favoritism shown in their department?	33	33	75	100
How many rarely (or never) get contradictory or conflicting orders?	50	50	86	87
How many rate their supervisor as being better than average as a person to work with?	17	20	92	100
How many are usually given reasons when changes are made in the way they are to do their work?	38	38	92	100
How many feel completely free to go to their boss for information or help when difficult problems come up in their work?	33	36	86	100
How many get the help they need when they do go to the boss for information or help on a difficult work problem?	28	33	75	100

[1] Lowest or Highest for each item. A number of *different departments* are represented in each column.

In a survey conducted a number of years ago, a situation was found that approached the above. In a relatively small warehousing operation, one third of the warehouse men had elected not to join the union. All men had the same general supervision and did the same kind of work. (1) On the majority of the survey items the responses

of the two groups of men were the same or differed very little (2) The union members expressed *greater satisfaction* on two topics *pay* and physical working con ditions and surroundings (3) The union members, however, felt slightly less secure on the job, felt less free to discuss prob lems frankly with the manager, criticized supervision more, were less satisfied with the company's promotion policies, and were not as well informed regarding com pany policies and operations

Groups such as the above are difficult to find But the proposed question can be answered partially by reviewing the survey results in companies in which some em ployees belong to a union and some do not A number of studies have been made where it was possible to segregate union members, working under the same general conditions (other than being subject to a union contract), in the same location with non union employees, but not under the same direct supervision In such studies the major *significant* difference found in the opinions expressed by union and non union employees has been in answer to questions about *pay* The former have ex pressed greater satisfaction with pay, both in terms of the amount received and in terms of comparing their pay with the pay for the same sort of work in other com panies

Though there are varying differences from company to company between union and non union employees on specific topics, it may be of interest to review the results from one survey which provides compari sons which seem representative of the type of findings observed

In this company about two fifths of the employees work under union contracts All but a few of the employees work in the same building All are subject to the same company policies and practices, and the same personnel policies except for a few modifications defined in the contracts (1) There were *very few* significant differences when the attitudes of union employees and non union employees were compared (2) The one outstanding and major difference was found in responses to questions about *pay*

Expressed dissatisfaction with PAY, de fined as the amount of money you make for the kind of job you have

Union members	30%
Non union employees	52%

Stated that their pay is as high as, or higher than the pay for the same sort of work in other companies

Union members	71%
Non union employees	42%

Except for differences on two topics re- lated directly to provisions in the union contracts, the other significant differences are summarized in the following state- ments (1) Union employees rated some of the company s benefit plans, especially the vacation policy, higher than did the non union employees (2) Non union em- ployees, as a group, felt more definitely that they "belonged, that they were a part of the organization (3) Union members were not kept as well informed about gen- eral company plans and policies, business conditions, new developments, reasons for changes, etc

Representative of items on which some, but not significant, differences were found are

	Proportion expressing opinion summarized at left	
	Union Members	*Non union Employees*
When I do some unusually good work, I usually or always, get recognition or praise for it	38%	45%
For getting ahead in this company, personal friendships with top executives count no more than is usual in companies of this size	79	73
In talking with my supervisor about my job, I could feel completely free to say exactly how I feel and to speak up about any complaints or problems that I might have	54	46

Representative of items on which very small (or no) differences were found are the following

	Proportion expressing opinon summarized at left	
	Union Members	*Non union Employees*
There are very few, or no other companies in which I would rather work at the same pay if I could get a job for which I feel equally qualified	80	79
I am very proud of the fact that I work for this company	61	59
I am getting a square deal from the company on my job here in most ways, or in every way	81	79
I can depend very well, or completely, on promises or statements made by my supervisor	57	56
I am rarely, or never, bawled out" or criticized when I do not deserve it	75	76
I am rarely, or never 'bawled out' or criticized in front of other employees	77	77
If I were to give my supervisor a good idea for a new or better way of doing a job I probably, or surely would get credit for it	82	80
When there is a better job vacant, the best qualified person usually, or always gets promoted to it	53	55

Output Rates among Chocolate Dippers *

HAROLD F ROTHE

There is a growing recognition among psychologists of the difficulties of obtaining adequate criteria in industrial work An increasing amount of space is being de voted to this topic in recent books and con ferences These difficulties are as true of objective performance data as of subjective criteria such as mer t ratings In a recent paper, Einar Hardin made particularly clear the difficulties involved in obtaining an adequate performance criterion [1]

Background of the Problem

In a previous paper the writer presented data on the output of butter wrappers, showing that a large number of daily work curves (based on production during each 15 minute period) must be obtained before a stable or consistent curve for an operator could be obtained Daily work curves for individuals had but little consistency from day to day (1) These butter wrappers showed an enormous range of productivity from one 15 minute period to the next, so

* Reprinted from *Journal of Applied Psychology* Vol 35, No 2 April 1951

[1] A paper presented at the Sixth Annual Conference on Industrial Relations Research held at the University of Minnesota, May 9 and 10, 1950

great that the variations within one operator s performance over a 5 day period were greater than the average variation from one operator to another over the same period (2)

In another paper on machine operators it was shown that the operators had relatively little consistency from one 2 week period to the next, these correlations being 57 and 68 in that study (3)

On the other hand, Tiffin has described a week to week consistency of 96 for hosiery loopers (4) It is probably noteworthy that the hosiery loopers were paid on an incentive system, and the butter wrappers and machine operators were not It is also probable that other factors such as shortages of material, good or poor scheduling, amount of work ahead of each operator, validity of recorded data, and even more subtle factors affect the consistency of the operators from period to period It is certainly clear that a great deal more detailed data must be obtained, together with a complete description of the setting in which the study took place, before industrial psychologists can reach sound conclusions regarding output data

The purpose of the present paper is to present another study, and a description of the setting of that study, as another step in the development of an understanding of output rates

Background of the Study

The data for this study were taken from the official books of the Fannie May Candy Company of Chicago, Illinois[2] They cover the period December 16, 1949 through April 15, 1950 They refer to regular, experienced female employees, working at their regular jobs at their usual workplaces This job is described in the USES Dictionary of Occupational Titles, code number 6 05 312

The Fannie May Candy Company is characterized by friendly employee management relations There has long been in effect an annual bonus system, a pension plan, a hospital plan, and similar aspects of progressive employee relations There is an employee representative committee that represents the employees Further, each employee feels free to walk into the General Manager s office at any time

There is an incentive system that has been in effect for many years and the employees are paid a 1 to 1 ratio for performance over standard It should be noted particularly in the data that follow here that this incentive system is an *effective* one, that is, the employees actually *do* regularly produce more than 100 per cent of performance, and get paid accordingly Finally, this job is completely controlled by the operator, there are no moving belts or machinery that govern the pace of the hand dippers

Since this study was limited to experienced hand dippers, and does not include persons on other jobs, part time dippers, or relatively inexperienced dippers or night shift dippers, the sample is small, being only 18 girls No data were included for any girl who did not work at least one full day on dipping in any one week, and, hence, the size of the sample dropped occasionally to a smaller number, but on no occasion were there less than 15 girls included in the data for any one week

TABLE 20 1

Weekly Average Output (Per Cent Performance of Standard) for Group of Chocolate Dippers

Week Ending		*Per Cent Performance*	*Number of Girls Weekly*
December	31	128 4	16
January	7	127 5	18
	14	129 5	18
	21	128 8	17
	28	129 0	17
February	4	126 9	18
	11	131 5	17
	18	127 5	15
	25	127 9	16
March	4	126 7	15
	11	127 2	17
	18	127 2	17
	25	132 7	16
April	1	134 2	16
	8	128 4	16
	15	126 0	16

[2] The writer wishes to thank Mr Harry H Simpson Vice President of the Fannie May Candy Company, for permission to publish these data

Data

Analysis of the trend of average weekly performance of the chocolate dippers shows that there was no consistent long term upward or downward trend in pro ductivity during this period These data are shown in Table 20 1 Thus, it may be concluded that the period of time studied here was an essentially normal ' one

The correlation of each operator s per formance for one week with her per formance for the following week was determined by the rank difference method The distribution of the obtained Rho s is shown in Table 20 2 The median inter

TABLE 20 2

Frequency Distribution of Rho s between Successive Weeks Output Individual Performance for Group of Chocolate Dippers

Rho	*Frequency*
96–1 00	2
91– 95	2
86– 90	3
81– 85	2
76– 80	2
71– 75	2
66– 70	0
61– 65	1
56– 60	1
Median Rho = 85	

weekly correlation is 85 Thus each girl was quite consistent in her performance from week to week

The greatest and least amount of pro ductivity for each girl for any one of the 16 weekly periods is shown in Table 20 3, together with the ratio of best to worst performance for each girl The median performance ratio of best to worst for these girls is 1 2 Thus, girl A produced a high of 135 per cent for one of these 16 weeks, and her lowest weekly average production during this time was 116 per cent, giving her a best/worst, or intra individual ratio of 1 16

The ratio of best operator to worst op erator, for each week, is shown in Table 20 4 During the week ending December 31, 1949, one girl produced 147 per cent

TABLE 20 3

Highest and Lowest Average Weekly Per formances, and Their Ratios for In dividual Chocolate Dippers During 16 Week Period

Girl	*Highest Weekly Average*	*Lowest Weekly Average*	*Ratio of Highest to Lowest*
A	135	116	1 16
B	123	95	1 29
C	142	120	1 18
D	175	144	1 22
E	134	108	1 24
F	114	104	1 10
G	131	103	1 27
H	149	128	1 16
I	146	132	1 11
J	136	125	1 09
K	138	108	1 28
L	150	136	1 10
M	124	114	1 09
N	145	117	1 24
O	140	126	1 11
P	147	103	1 43
Q	140	113	1 24
R	133	120	1 11
S	152	127	1 20
Median intra individual ratio = 1 18			

TABLE 20 4

Highest and Lowest Average Individual Weekly Performances, and Their Ratios, for Group of Chocolate Dippers During 16 Week Period

Week Ending	*Highest Girl's Average*	*Lowest Girl s Average*	*Ratio of Highest to Lowest*
December 31	147	105	1 40
January 7	149	101	1 48
14	150	102	1 47
21	152	107	1 42
28	150	99	1 52
February 4	149	99	1 51
11	175	101	1 73
18	144	98	1 47
25	157	95	1 65
March 4	155	98	1 58
11	156	110	1 42
18	156	106	1 47
25	150	113	1 33
April 1	159	104	1 53
8	157	101	1 55
15	148	103	1 44
Median inter-individual ratio = 1 475			

and another gırl produced 105 per cent, for an ınter ındıvıdual ratıo of 1 40

From Tables 20 3 and 20 4 ıt can be seen that the average ratıo of the range of ınter ındıvıdual performance ıs greater than the average ratıo of the range of ıntra ındıvıdual performance

Discussion

The present wrıter has prevıously stated two tentatıve hypotheses relatıng to the effectıveness of ıncentıves One was that the ıncentıves to work may be consıdered ıneffectıve when the ratıo of the range of ıntra ındıvıdual dıfferences ıs greater than the ratıo of the range of ınter ındıvıdual dıfferences' (2, p 326) It ıs quıte ap parent that here, where we may safely conclude that the varıous ıncentıves were effectıve, that thıs hypothesıs ıs supported by these data

The second hypothesıs was that "ıf the ıntercorrelatıon of group output rates for two perıods closely related ın tıme ıs less than 80 the ıncentıvatıon ıs not hıghly effectıve, whıle ıntercorrelatıon hıgher than 90 ındıcates effectıve ıncentıvatıon ' (3, p 488) Thıs hypothesıs ıs not supported by thıs study, although the error ıs ın amount, not ın dırectıon It ıs, of course, rısky busıness makıng a hypothesıs wıth as lıttle data as exıst ın thıs field On the other hand, hypotheses may stımulate someone else to do research (and to dıs prove them)

On the strength of other data whıch have not been publıshed, the wrıter belıeves that there ıs a relatıonshıp between consıstency of data and effectıveness of ıncentıves He also belıeves that the ıncentıvatıon of the machıne operators (prevıously descrıbed) was actually faırly hıgh (3), even though there was no ıncentıve system there at the tıme That ıs, hıgh ıncentıvatıon probably exısted there because of an ıntangıble attı tude and ın spıte of the lack of an engı neered ıncentıve system as such at the tıme of that study

Hence, the second hypothesıs ıs herewıth changed to read, "ıf the ıntercorrelatıon of output rates for two perıods closely re lated ın tıme ıs less than 50 the ıncentı vatıon ıs not hıghly effectıve, whıle ıntercorrelatıon hıgher than 80 ındıcates very effectıve ıncentıvatıon "

Thıs hypothesıs ıs of course tentatıve It ıs also lıkely that these coefficıents wıll vary wıth the length of the perıods of tıme correlated Presumably lower correlatıons wıll be found ıf short perıods of tıme are correlated than ıf long perıods are corre lated

Another poınt ıs also ımportant here It ıs apparent that the wrıter ıs attemptıng to demonstrate somethıng other than 're lıabılıty by proposıng these hypotheses That ıs, ıt appears that the consıstency of output, from tıme to tıme, may reflect a real phenomenon and permıt some predıc tıon (such as an estımate of the effectıve-ness of the ıncentıves ın a sıtuatıon) and not merely reflect the ' relıabılıty or un relıabılıty" of the crıterıon The concept of relıabılıty' has been taken over from the realm of testıng where ıt ıs related prımarıly to errors of measurement Such errors undoubtedly exıst ın work such as ıs descrıbed ın thıs artıcle But ıt ıs also very lıkely that there ıs a ' consıstency or 'ınconsıstency" of ındustrıal output data that ıs a real phenomenon and not merely an error of measurement[3]

Summary

Analysıs of the output rates of 18 choc olate dıppers for a 16 week perıod, under condıtıons of excellent ıncentıvatıon, re veals an average week to week consıstency coefficıent of 85 and an average ratıo of the range of ınter ındıvıdual performance greater than the average ratıo of the range of ıntra ındıvıdual performance Both these fındıngs tend to support some hypotheses prevıously stated by the wrıter, although wıth changes ın one of the hypotheses

The probabılıty that coefficıents of con sıstency of output rates, from tıme to tıme, reflect some real phenomenon or co varıate, and are not merely ındıcatıve of errors of measurement (relıabılıty) ıs dıscussed

References

1 Rothe, H F Output Rates Among

[3] Thıs poınt has also been expounded by Eınar Hardın, prevıously cıted It ıs sın cerely hoped that he wıll soon publısh hıs paper on the subject

Butter Wrappers I" *Journal of Applied Psychology* 1946 Vol 30, 199–211
2 Rothe, H F, Output Rates Among Butter Wrappers I,' *Journal of Applied Psychology* 1946, Vol 30, 320–327
3 Rothe, H F, 'Output Rates Among Machine Operators I, *Journal of Applied Psychology* 1947, Vol 31 484–489
4 Tiffin J, *Industrial Psychology* New York Prentice Hall, 1942

Chapter V

LABOR MANAGEMENT RELATIONS

The extent and duration of the labor-management controversy is ample evidence of the conflict existing between the groups involved Laws, as such, do not usually solve the problem, sometimes the mere passage of a law seems to intensify the conflict

Although a basis for disagreement between management and labor is likely to continue, a more mature manner of easing the situation can be evolved To be sure, most of the skull-cracking days have passed, but we must still persevere before we can be regarded as handling this problem in normal adult fashion

Psychologists can contribute greatly toward a happier solution to this problem Research related to the roots of the problem will add to needed knowledge—recognition that industrial conflict stems from conflicting motives of individuals and groups is of paramount importance Exposing individuals to experimentation in this field is difficult to justify, simply because people would be hurt Other avenues of research, however, are available The investigations of Porter, Eckerman, and Weschler, it is hoped, are merely forerunners of many more studies that will contribute to a more peaceful solution of this intensely important problem

Porter has analyzed arbitrations of certain industrial disputes, and concludes that secondary motivations of status and group loyalty are of real significance

Eckerman has demonstrated that data concerning grievances and those responsible for making the complaint are subject to statistical analysis, and believes that such information can lead to a better understanding of such problems

Weschler has used a technique known as "error choice" to determine bias in the direction toward labor or management In another study, he attempts an analysis of characteristics of "good" and "poor" mediators Most interestingly, through the "error choice" method, he establishes that not all mediators are neutral, and that this factor is related to a differentiation of "good" and "poor" mediation

The Arbitration of Industrial Disputes Arising from Disciplinary Action *

J M PORTER, JR

The present paper reports on an initial study of industrial disputes arising from disciplinary action which have been taken to arbitration Our interest in this area arises from the question as to the effect arbitration of such disputes has upon management's effectiveness in maintaining discipline within the plant We have not as yet found the answer to such questions but our study of nearly 200 arbitration awards has revealed some pertinent information about the behavior of the parties involved

Analysis of Arbitration Awards

The material selected for initial study consisted of 197 arbitration awards in which the issue was the equity of the discipline (suspension or discharge for the greater part) imposed upon individual employees, for behavior which management deemed detrimental to the effective operation of the plant These awards were all those involving the arbitration of disciplinary disputes reported by a leading industrial relations reporting service during the years 1946 and 1947

A wide variety of production and service industries were represented In only a few cases did the discipline administered apply to more than a single employee Sixty-eight different individuals served as arbitrators and the median number of awards per arbitrator was slightly more than one

In 74 cases, which represent 38 per cent of the total number studied, the arbitrator's award sustained the disciplinary action taken by management In the remainder of the cases studied, 121, which was 62 per cent of the total, the effect of the arbitrator s award was to either revoke or modify the discipline imposed by management In this latter group of cases, the effect of the award was to completely revoke management's action in 49 per cent of the instances and to modify (i e, reduce in severity) the discipline in 51 per cent of the instances

Where the original discipline had taken the form of a suspension (24 cases) the arbitrator s decision sustained the action taken in two thirds of the instances Where the original discipline imposed had been the discharge of the employee (170 cases) the arbitrator's award sustained the action in 34 per cent of the cases Within the limits of the cases studied, there appears to be a definite tendency for the arbitrator to modify the discipline imposed by management However, when suspension rather than discharge is involved, the awards sustain management's actions two to one

Our figures show that discharge is the form of discipline most frequently resorted to by management It must be borne in mind, however, that our data were gathered from disputes which had been taken to arbitration and such findings may merely mean that unions are more apt to press discharge cases to arbitration than lesser forms of discipline

We next attempted to formulate the categories of employee behavior which evoked disciplinary action We first listed all the forms of behavior cited by management at the arbitration hearing in substantiation of its actions In some cases more than one form of behavior on the part of the disciplined employee was cited Where this was the case and the company s argument dwelt at any length upon more than one kind of behavior as contributing to their decision, the cases have been classified in as many categories as were appropriate Where the bulk of the company s argument was confined to the fact that the em-

* Reprinted from Proceedings of Second Annual Meeting, Industrial Relations Research Association, December, 1949

TABLE 21 1

Violation of Shop Rules Cited 59 (28%)			Incompetence and/or Inefficiency Cited 41 (20%)			Insubordination Cited 55 (27%)			Violation of Contract Cited 42 (21%)		
Mgt Sust 45%	Discip Mitig 55%		Mgt Sust 32%	Discip Mitig 68%		Mgt Sust 40%	Discip Mitig 60%		Mgt Sust 31%	Discip Mitig 69%	
	Revoked 39%	Pen Reduced 61%		Revoked 68%	Pen Reduced 32%		Revoked 51%	Pen Reduced 49%		Revoked 37%	Pen Reduced 63%

ployee's behavior had been of one kind, the case was of course classified under a single heading We find that four categories of behavior are sufficient to classify all but a very minor number of cases

Attention is now called to Table 21 1

Violation of shop rules was cited 59 times This is 28 per cent of the total citations Into this category we have classified such forms of behavior as intoxication on the job, tardiness, fighting with co workers, gambling, dishonesty, excessive absenteeism, absence without proper notice, and failure to report for work when scheduled

When the reasons cited by the company for the discipline imposed were violations of shop rules, the arbitrator's award sustained management in 45 per cent of the cases and mitigated the discipline imposed in 55 per cent of the cases In dealing with those cases in which the effect of his award was to mitigate the discipline imposed, the arbitrator completely revoked management's action in 39 per cent of the cases and acted to reduce the severity of management's penalty, in 61 per cent of the instances

Incompetence and/or inefficiency was alleged by the company in 41 cases as the reason for discipline This represents 20 per cent of the total citations Into this category we have classified behavior described as an uncooperative attitude, carelessness, negligence, conducting personal business on company time, and the general statement that the employee was incompetent or inefficient on the job, or both

In such cases the effect of the arbitrator's decision was to sustain management's action, i e, find that sufficient cause existed, in 32 per cent of the cases In the remaining 68 per cent of the cases, in which the effect of his award was to mitigate the discipline, the award had the effect of revoking it completely in 68 per cent of the cases and of finding that cause for discipline existed but that the action taken by management was too severe in the light of the apparent facts in 32 per cent of the instances

Insubordination was cited by the company in support of the discipline administered in 55 instances This represents 27 per cent of the total citations In addition to the general statement that the disciplined employee's behavior had been insubordinate, such specific acts were classified as representing insubordination as fighting with the supervisor, friction with the foreman, use of profanity in arguing with the boss, and refusing proper work assignments

When insubordination was cited as the reason for the discipline administered, the effect of the arbitrator's award was to sustain management's action in 40 per cent of the instances and to mitigate it in 60 per cent of the cases In the latter instances, i e, the cases in which the arbitrator mitigated the discipline imposed, the arbitrator completely revoked the discipline almost as frequently as he indicated that he felt some penalty was merited but that the penalty assessed by management was too severe

Violation of the labor management agreement (the contract) was given as the reason for discipline in 42 cases This represents 21 per cent of the total number of citations Such behavior as interference with the direction of the working force, engaging in work stoppages and slowdowns, coercion and solicitation of workers to join the union on company time and property, and refusing to follow the grievance procedure as set forth in the agreement were classified as violations of the contract

When violations of the labor management agreement were cited as meriting the discipline imposed, the effect of the arbitrator's award was to sustain management in 31 per cent of the instances and to mitigate management's action in 69 per cent of the cases In those cases where the effect of the arbitrator's decision was to mitigate the discipline imposed, the action of management was completely revoked in 37 per cent of the cases, in 63 per cent of the instances the arbitrator indicated that he felt some discipline was merited but that imposed by the company had been too severe

Only 4 per cent of the cases cited grounds for discipline which could not be classified under one or the other of the four categories just described

Summary of Statistics

In summary, our study has indicated that arbitration sustains management's imposition of discipline in approximately 40 per cent of the cases and mitigates the discipline imposed in approximately 60 per cent of the cases. The behavior held most frequently by management to merit discipline was (1) Violation of Shop Rules which accounted for 28 per cent of the cases (2) Incompetence and/or Inefficiency which accounted for 20 per cent of the cases (3) Alleged Insubordination which accounted for 27 per cent of the cases (4) Violation of the Labor Management Agreement which accounted for 21 per cent of the cases.

Violation of shop rules and insubordination are cited as reasons for discipline with slightly greater frequency than incompetence and/or inefficiency and violation of the labor management agreement.

Our analysis shows that the arbitrator's award mitigated the discipline imposed most frequently when the behavior cited as meriting the discipline was incompetence and/or inefficiency and violation of the labor management agreement. Discipline for violation of shop rules is mitigated least frequently.

Where the effect of the arbitrator's award was to mitigate the discipline imposed, arbitrators, when they have felt discipline was merited, have been more inclined to substitute their opinion for that of management in determining the discipline merited when violation of shop rules and violation of contract were alleged than they have been when incompetence and/or inefficiency and insubordination were alleged. When insubordination was alleged, and the arbitrator's award mitigated management's action, the arbitrator completely revoked management's action about as frequently as he decided that while a penalty was merited, that fixed by management was too severe. When incompetence and inefficiency were alleged, the arbitrator's award completely revoked the discipline imposed more than twice as frequently as deciding that though a penalty was merited, that fixed by management was too severe.

Motivations Revealed by Case Analyses

The study of the arbitration of industrial disputes arising from disciplinary action affords an opportunity for studying significant forms of social activity. The issues presented for adjudication permit the study of human motivations, interaction, and conflict in an atmosphere relatively free from appeals to previously established doctrines of *stare decisis* and similar forms of precedent.

It may be expected that the study of various aspects of industrial arbitration, particularly of disputes arising from disciplinary action, will reveal data with respect to human motivations which will supplement that obtained through attitude and morale surveys. In our opinion, insights gained through the study of the activities of the groups and the issues involved in arbitration have a validity which insights gained through the other methods of study lack. Arbitration is an activity in which the parties do not engage unless they feel strongly about the issues involved, and it is reached only after prior efforts by the parties themselves to settle the dispute.

We have also been interested in such insights as might be achieved into the thought processes of the arbitrator himself. This individual is coming to play an increasingly significant role in industrial relations. As the fact of their responsibility to the public has been increasingly impressed upon labor and management groups, we find increased resort to arbitration for the peaceful resolution of conflicts which the parties are not able to resolve themselves.

The behavior evoking discipline which we have classified as a violation of shop rules or that indicating incompetency and/or inefficiency apparently represents a conflict between the individual's personality characteristics and conditions established by management in order to conduct the business of the company in what is deemed an efficient and effective manner. In our experience, open feelings of hostility are not usually present in such instances and the dispute is generally a question of whether or not a violation of shop rules

occurred or whether or not incompetency or inefficiency was present And, if so, has the discipline under consideration been applied consistently throughout the plant in the past? The problem posed for the arbitrator in such cases is generally a question of the determination of the fact of the violation or incompetence and management s consistency in the application of the rules or standards which have been set up

On the other hand, cases identified as involving insubordination and contract violation are frequently situations of open hostility and the thinking of both parties is emotionally prejudiced thereby Where an employee is disciplined for either of these causes, the motivation behind the discipline may be at least partially a reaction to implied loss of status on the part of management—a threat to the authority and prestige of the management man This is particularly true at the lower levels of management where most disciplinary action initiates

This hypothesis is indicated by the fact that violations of shop rules and incompetence and inefficiency on the job seem to elicit disciplinary action more frequently when associated with behavior on the employee s part which the supervisor or foreman interprets as insubordinate Thus there is frequently an over reaction, by first line management particularly, when insubordination or violation of contract—matters of status—are at stake The discipline applied need only meet the requirements of management s responsibility for efficient production in the plant, but, in fact, it tends to exceed this need and becomes an action mainly of vindication of status and exercise of authority

The need for status, frequently coupled with adherence to feelings of group loyalty, also seems to motivate the union's activities in arbitration hearings Many cases of violation of shop rules and discipline for incompetence are brought to arbitration by the union even though the facts seem clear that the employee has been guilty of the conduct alleged and no discrimination in the application of the rules or standards by management is evident The union leadership fights the disciplinary action simply because its status with its members has been challenged and group loyalty tested The motive of status protection is even more clearly defined in the union s defense of alleged acts of insubordination or contract violation

These secondary motivations, status and group loyalty, play an important role in the process of arbitration of disciplinary action By the language of the labor management agreement the parties are concerned only with the question of the justice of the discipline imposed Has management's action been taken for "proper or just cause"? Yet the number of instances which occur wherein the union challenges non discriminatory discipline for patent violation of shop rules, obvious incompetence or unmistakable insubordination, and contract violation, and wherein management resorts to extreme disciplinary action in the case of mild insubordination or violation of contract, seem to indicate that the equity considerations outlined in the Agreement are not the sole motivation A clear understanding of these underlying motivations would, we think, not only eliminate a substantial number of disciplinary disputes but significantly aid plant morale as well

Also of value in the application of this understanding of motivation might be more frequent resort by management to suspension rather than discharge If one views the objective of disciplinary action as the improvement of behavior, then it is clear that insofar as the individual disciplined is concerned, any value in terms of reformed behavior is lost to the company when the man is discharged Also lost is the company s investment in the training given that man Moreover undue discharge has a negative effect upon the other employees Thus those instances which give rise to the opinion that discharge is invoked rather than suspension because of management s over reaction to a threat to status, would be eliminated and the relations between labor and management would be expected to be benefited thereby

These unspoken secondary motivations of status and group loyalty impinge upon the behavior of the arbitrator as well The cases studied demonstrate this premise in

two extremes of approach employed by the arbitrator One criterion of judgment is to simply determine whether management has proper cause for discipline and had been consistent in the past in its application and, if so, sustain management s action irrespective of its severity The other standard of judgment adopted by some arbitrators is, after determination of proper cause and non discrimination, to independently evaluate the discipline in terms of what the arbitrator considers merited and so award Under this second point of view, the arbitrator frequently modifies the company's disciplinary action even though a finding of proper cause and good faith on management s part has been made Our study indicates that this substitution of the arbitrator s judgment for that of management properly exercised, is more frequent in instances of management discipline for violation of shop rules and contract violation

CONCLUSION

We may tentatively conclude from this preliminary study that secondary motivations are of real significance in the arbitration of industrial disputes which arise from disciplinary action That both parties are significantly influenced by considerations of status, and that the union is additionally influenced by consideration of group loyalty also seems apparent The operation of these needs interferes with sound function of the disciplinary process and the arbitration of disputes arising therefrom The arbitrator, in turn, often has a tendency to go beyond the authority contractually vested in him to ascertain proper cause and non discrimination, and substitutes his judgment for that of management in determining the appropriateness of the discipline meted out We believe that further investigation would shed worthwhile light on behavior in this field and aid in the understanding of the disciplinary and arbitration process

An Analysis of Grievances and Aggrieved Employees in a Machine Shop and Foundry *

ARTHUR C ECKERMAN

This article is based on the author's dissertation of the same title submitted to the Faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, February, 1948 The dissertation was directed by Dr Joseph Tiffin

The person working on the labor relations front in industry knows how slowly progress is being made in bringing labor and management to a closer understanding of their mutual problems, problems which are inevitably reflected in the endless grievances that must be processed Any program which simply proposes a different method of handling grievances is only dealing with symptoms, the real causes underlying the complaints of labor will lie untouched

Various discussions concerning the problems of present day labor relations led to the research represented by this paper The hypothesis was developed that a statistical analysis of grievances might indicate significant differences existing between employees having grievances and non aggrieved employees

A large Midwestern plant allowed this study to be made of its grievances and aggrieved employees The name of the city, state, and company is withheld, and the nature of the company's products is unimportant to the research Two unions had

* Reprinted from *Journal of Applied Psychology*, Vol 32, No 3, June 1948

contracts with the plant, a machine shop union and a foundry union

Grievances at the plant were divided into two classes, oral and written As no record was kept of grievances in the first and second steps it was impossible to get an estimation of the number and the nature of these grievances After the grievance had been reduced to writing in the third step a complete and accurate file was kept on it regardless of its disposition Therefore, in this study of the grievances of the plant only those grievances were used that had reached the third step with subsequent reduction to writing

This situation was fortunate from two standpoints First, disregarding steps one and two probably reduced the size of the study considerably and simplified it Secondly, by not using the first two steps of the grievance procedure the results are probably more valid from the standpoint of being an accurate description of real labor problems in the plant Grievances in steps one and two may perhaps more correctly be described as complaints of employees Only when these complaints are found to be real differences of opinion between the thinking and the program of the union, on one hand, and the thinking and the policies of the company, on the other, may they be considered true grievances At this stage they are formalized by writing and are taken out of the hands of operating supervision and operating union officials alike to become matters of genuine concern of the managements of both the union and the company

The research was undertaken with no thesis in mind, there was no thought of proving any preconceived opinions, either unionwise or companywise The only hypothesis was that if significant differences exist between aggrieved and non aggrieved employees, this type of research might identify and describe those differences

It is hoped that this research will be of some help to American labor and industry in their ceaseless striving to arrive at a better understanding of their mutual problems

Procedure

A survey of the grievance files of the plant revealed a source of data complete in detail for each grievance and in chronological order The personnel records of the plant were also in excellent order and quite complete

A work sheet was made up which contained two sets of data, (a) the pertinent facts of nine items of the grievances, and (b) all available information concerning the aggrieved, which consisted of 75 items From the very complete grievance and personnel records of the company 1067 work sheets were filled in which represented 766 separate grievances of 327 employees A number of employees, mostly union officials, each filed more than one grievance The average number of grievances per employee of the group having grievances, was 2 3 The first grievance filed by an employee was designated as an 'initial grievance" Table 22 1 shows the distribution of grievances and grievers

In the foundry agency 223 initial grievances were found and 104 in the machine shop Work sheets were made up for two control groups consisting of 201 foundry employees, selected at random from the personnel files, who had not filed a grievance and 100 machine shop non aggrieved

TABLE 22 1

Number of Grievers and Grievances

	Foundry	*Machine Shop*	*Total*
Grievers	223	104	327
Initial	150	92	242
Other	73	12	85
Grievances	644	122	766
Initial	223	104	327
Other	421	18	439

employees also selected at random from the personnel files

Items on the work sheets were coded, and the resulting information was punched on I B M cards Two sets of data were tabulated from the punched cards, data on the grievances and data on the grievers and the control groups The foundry data were separated from that of the machine shop Each bargaining agency then had two sets of data, grievance information and personnel information The grievance data had two divisions, that of (a) initial grievances and (b) other grievances The personnel data also had two divisions, (a) aggrieved and (b) non aggrieved employees

A statistical analysis was made of the results of the tabulation The figures were expressed in either per cents or medians A number of items such as vacation, yearly income, and others were calculated only for a twelve month period, the calendar year of 1946 All figures are comparable, having been equated to take care of variables introduced by a general wage raise

The difference between each of the respective groupings of grievance and personnel data was computed The standard error of each difference was also computed Fisher s *t* statistic was computed by dividing the difference by the standard error of the difference An entry from Fisher s table was then obtained for each *t* value and the probability *P* determined

Each *t* value is indicative of a level of significance which may be interpreted as the probability that a difference as large as the obtained difference could have occurred if the samples were drawn from the same population, or to put it another way, as the probability that the difference could have occurred by chance alone

If a difference as large as the obtained difference could have occurred as frequently as 5 times in 100 among pairs of samples drawn from the same population, the difference is considered significant at the 5 per cent level Since chance alone could account for a difference as large as the obtained difference only 5 times in 100, the null hypothesis that the true difference is zero can be rejected A difference significant at the 5 per cent level or lower is marked on the tables with a double dagger

If a difference as large as the obtained difference could have occurred as frequently as 10 times in 100 among pairs of samples drawn from the same population, the difference is considered significant at the 10 per cent level A difference significant at the 10 per cent level is marked on the tables with a dagger If a difference as large as the obtained difference could have occurred as frequently as 20 times in 100 among pairs of samples drawn from the same population, the difference is considered significant at the 20 per cent level A difference significant at the 20 per cent level is marked on the tables by one asterisk

These asterisk markings of the three levels of significance are made to facilitate a rapid identification of the most probably significant items

The results of the analysis of data concerning grievances is expressed by comparing the respective standings of union members and union officials on each item "Initial grievances are those filed by union members on their own behalf, "other grievances' are those filed by union officials, usually for a cause furthering the union s program or the operation of the agreement

The results of the analysis of all data concerning aggrieved employees are expressed by comparing the respective standings on each item of the grievers and a control group of employees having no grievances

Only those items were used in comparing the respective groups where the number of cases involved was large enough for statistical handling Differences between the groups of grievers and non grievers that are probably most significant are those in the *t* column of the Tables which are marked with three asterisks Values of *t* which are marked with two asterisks might be considered significant, but as these values decrease they are to be interpreted with increasing caution as chance factors are more apt to be responsible for the difference as *t* values become smaller

TABLE 22 2

Grievance Data of Foundry and Machine Shop Employees †

	Foundry				*Machine Shop*			
	Initial Griev ances	Other Griev-ances	Differ-ences	*t*	Initial Griev ances	Other Griev-ances	Differ ences	*t*
Contract Not Referred to	72 6	83 1	−10 5	1 39*	50 0	55 6	− 5 6	30
Classification of Grievances								
Job and Work	24 8	33 6	− 8 8	1 25	14 4	27 8	−13 4	61
Pay and Wages	29 3	25 0	+ 4 3	61	49 0	44 4	+ 4 6	24
Seniority	15 3	4 0	+11 3	1 45*	20 2	11 1	+ 9 1	38
Steps Grievance Settled in								
Third Step	73 5	69 8	+ 3 7	85	51 9	50 0	+ 1 9	11
Fourth Step	22 4	20 9	+ 1 5	20	28 8	44 4	−15 6	80
Fifth Step	3 6	8 8	− 5 2	64	18 3	5 6	+12 7	52
Disposition of Grievance								
Granted by Company	35 4	35 4	0 0	00	35 6	66 7	−31 1	2 04‡
Denied by Company	61 4	62 2	− 0 8	16	60 6	33 3	+27 3	1 35*
Dropped by Union	2 7	2 1	+ 0 6	07	3 8	0 0	+ 3 8	40

* Significant at the 20 per cent confidence level
† Grievance data in this table are expressed in per cents
‡ Significant at the 5 per cent confidence level

Results

Grievance data of foundry and machine shop employees Nine items concerning grievances were available Several of the items such as the classification of grievances and the disposition of grievances, had subdivisions Only those items are shown in Table 22 2 which had large enough numbers of cases to justify the computation of differences

a Union officials, as reflected in the grievances they file, do not refer to the contract in the wording of their grievances as often as do union members This difference between the two groups is probably greater in the foundry unit than that of the machine shop

b Relative to the nature of grievances more grievances appeared to be filed by union officials concerning work and jobs than by union members, particularly in the foundry unit, but the differences are not significant Union members file more grievances concerning pay and wages than do union officials However, of the grievances concerning job and work or pay and wages union members' grievances show a larger percentage of the latter This is particularly true in the machine shop where the ratio is approximately one to three Union members in the foundry file more grievances on seniority than do their officials The difference is not significant in the machine shop where the same relationship appears to hold A higher percentage of machine shop grievances of union members are concerned with seniority than among the foundry group, although union officials of the machine shop agency are more concerned with seniority problems, as reflected in their grievances, than are the foundry union officials Percentage figures on the other six items of the classification of grievances may be found in Table 22 3

c In the settlement of grievances no significant differences were found between the subject groups Grievances settled in the third step had values in the same direction, in favor of the union members, in both the foundry and the machine shop The number of grievances involved in the fifth step was too small to make a reliable comparison The sixth step, arbitration, also had too few cases, there being only four with which to make any comparisons

d The machine shop grievances of union officials are definitely granted by the company more often than those of union members There is no difference between these two groups in the foundry This would seem to indicate that union officials of the foundry unit are not as able in formulating and processing grievances as are the union officials in the machine shop The converse is likewise true in the machine shop unit more grievances of union members are denied by the company than are those of union officials It is indicated that in the machine shop, the grievances that are dropped by the union are those of its members and not those of its officials, but the difference is not significant In the foundry an equal number of grievances of both groups seem to be dropped by the union

Personal data of foundry and machine shop employees Seventeen items of personal data were available on the employee s personnel record These data were compared for the two groups, the grievers and the non grievers Table 22 4 gives the figures on the comparisons for these two groups in the foundry and in the machine shop An analysis of the results obtained, when the personal data of aggrieved employees of the foundry and the machine shop were compared with that of their corresponding control groups, showed several significant differences between the two groups, the grievers and the non grievers

a Little difference was found between grievers and non grievers in the foundry and machine shop in regard to education More of the machine shop grievers went farther than the eighth grade than did the non grievers

b It is indicated that foundry grievers are socially more stable in that fewer of them are single, more of them are married and more of them have children A greater number of the foundry grievers have children than do the machine shop grievers As the needs of a family group are greater than a family without children and as grievances filed were primarily for more money, it would seem to follow that in the

TABLE 22 3

Number of Grievances by Classification †

Grievances	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
Foundry, Initial Grievances	55	66	34	20	9	16	11	5	7	223
Foundry Other Grievances	15	51	21	6	0	3	4	3	1	104
Machine Shop Initial Grievances	142	105	17	15	7	4	52	45	34	421
Machine Shop, Other Grievances	5	8	2	0	0	0	2	1	0	18
Total Grievances	217	230	74	41	16	23	69	54	42	766

† Key to classification of grievances

1—Job and work
2—Pay and wages
3—Seniority
4—Promotion and transfer
5—Vacation
6—Discharge and reinstatement
7—Union business
8—Company business
9—Matters for collective bargaining or mutual agreement

TABLE 22 4

Personal Data of Foundry and Machine Shop Employees

	Foundry				*Machine Shop*			
	Grievers	Non Grievers	Differ ences	t	Grievers	Non Grievers	Differ ences	t
Age in Years (Mdn)	33 6	34 0	− 0 4	1 2	36 0	36 5	− 0 5	24
Height in Inches (Mdn)	69 5	69 2	+ 0 3	1 2	69 4	69 2	+ 0 2	37
Weight in Pounds (Mdn)	159 4	156 3	+ 3 1	1 2	158 1	156 8	+ 1 3	38
Birthplace (% Born in)								
City	19 1	21 4	− 2 3	28	13 5	20 4	− 6 9	53
State (Excl City)	23 2	14 0	+ 9 2	1 02	60 6	52 0	+ 8 6	92
North (Excl State)	7 7	6 5	+ 1 2	13	9 6	3 0	+ 6 6	49
South	49 4	57 7	− 8 3	6 38‡	16 3	24 5	− 8 2	5 01‡
Schooling (%)								
8th Grade and Under	45 7	52 1	− 6 4	92	25 3	37 0	−11 7	1 00
Over 8th Grade	54 3	47 9	+ 6 4	93	74 7	63 0	+11 7	1 49*
Marital Status (%)								
Single	6 7	22 4	−15 7	1 8†	14 4	23 0	− 8 6	68
Married	92 4	75 7	+16 7	4 2†	82 6	75 0	+ 7 6	1 2
Children	63 7	46 8	+16 9	2 6‡	56 7	50 0	+ 6 7	70
Other Dependents	13 9	20 4	− 6 5	74	11 5	9 0	+ 2 5	19
Non Veteran	81 6	68 9	+12 7	1 26	66 2	68 0	− 1 8	22
Application Form								
Jobs Listed (Mdn Number)	4 3	3 0	+ 1 3	6 6‡	5 7	4 3	+ 1 4	3 2‡
Total Time on (Mdn Months)	42 5	24 3	+18 2	7 0‡	66 5	37 2	+29 3	2 9‡
Employed at Time of Application (%)	65 5	51 3	+14 2	2 29‡	48 4	55 8	− 7 4	75

* Significant at the 20 per cent confidence level
† Significant at the 10 per cent confidence level
‡ Significant at the 5 per cent confidence level

two machine shop groups employees with families should have more grievances Differences run in the same direction as those in the foundry, but they are not significant

c On the application form of the company, applicants hired who later filed grievances had had more jobs and had worked longer than applicants who became non aggrieved employees Among the foundry group, more of the subsequent grievers had jobs when they applied for work at the plant than did the subsequent non aggrieved employees The reverse was true in the machine shop group where more of the non grievers had jobs at the time of application, but the differences here were not very pronounced Two reasons might be postulated why foundry workers who have grievances were looking for different jobs, (a) due to more of them having families they were in need of better paying jobs, (b) as a group they appear to be less stable socially than non grievers

d A greater per cent of employees born in the South were non grievers rather than grievers in both the foundry and machine shop A study of other places of birth revealed no significant differences between grievers and non grievers although all differences were in the same direction for both the foundry and machine shop

e No appreciable differences were found between the grievers and non grievers with respect to weight, height, or age This is true of both the machine shop and foundry groups although the differences found were in the same direction for both groups

Personnel records data of foundry and machine shop employees Of the numerous items taken from the personnel records, those in Table 22 5 proved to be most interesting

a More foundry grievers had personnel transactions than non aggrieved employees This indicates that foundry employees who are subject to personnel changes, regardless of the nature of the transaction, are most liable to be grievers

b Aggrieved employees have more total net service than do non-grievers The machine shop grievers have over two years more service than non grievers whereas the foundry grievers show approximately one year more service than foundry non grievers

c New employees who later became aggrieved for one reason or another started at about 18 cents an hour less than employees who did not have grievances This is true in both the foundry and machine shop However, it is indicated that at the time of the grievance, the aggrieved employees were slightly higher in hourly rate than non aggrieved employees, although the difference is not significant

d The total wage increase of the grievers from time of hire to the time of grievance was significantly higher than the wage increases received by the non aggrieved group for a corresponding period It is demonstrated by this study that both the union and the company were cooperating in erasing wage differences between workers the union in pointing out the differences and the company in equating them

e The grievers as a group were more subject to layoff than were the non grievers, particularly in the foundry However, when it came to temporary layoffs, over twice as many non-grievers took temporary layoffs as did the grievers This would seem to indicate that the plant supervision was well aware of problems it could make for itself, but on the other hand it has been noted that grievers have considerably more net service, hence more seniority, than do non-grievers

f In the matter of skill level of the respective groups, the semi skilled bracket of the foundry had the highest percentage of grievers in it The most grievers fall in the semi skilled bracket of the foundry and machine workers It is indicated that the semi skilled level of foundry employees is most liable to produce grievers

g There seemed to be little difference between the annual earnings of the two groups although it is indicated the grievers in the foundry earned slightly more money in the year of 1946 than did the non-grievers In comparing the annual wages of veterans and non veterans, it is indicated that in both the machine shop and foundry veterans who were non-grievers earned more per year than veterans who were grievers, particularly in the machine shop group The same relationship exists though

TABLE 22.5

Personnel Records Data of Foundry and Machine Shop Employees

	Foundry				Machine Shop			
	Grievers	Non-Grievers	Differ-ences	*t*	Grievers	Non-Grievers	Differ-ences	*t*
Total Net Service in Mo. (Mdn.)	39.2	27.8	+ 11.4	5.46‡	68.0	41.1	+ 26.9	5.62‡
Starting Rate (Mdn. ¢)	.62	.78	− .16	16.36‡	.60	.77	− .17	2.71‡
Rate at Time of Grievance (Mdn. $)	1.10	1.09	+ .01	.10	1.09	1.06	+ .03	.3
Total Wage Increase to Date of Grievance (Mdn. ¢)	.48	.30	+ .18	24.3‡	.48	.34	+ .14	3.85‡
12 Month Earnings (Mdn. $)	2881	2813	+ 68	1.42*	2742	2676	+ 66	.60
Veteran	2261	2439	−178	.99	2287	2601	−314	1.79†
Non-Veteran	3010	3038	− 28	.04	2907	2701	+206	.59
White Employees (%)	45.8	29.9	+ 15.9	2.06‡	99.2	98.0	+ 1.2	.13
Skill Level of Employees (%)								
Skilled	11.7	10.9	+ 0.8	.09	27.9	16.0	+ 11.9	.96
Semiskilled	67.2	51.2	+ 16.0	2.56‡	68.3	65.0	+ 3.3	.41
Unskilled	21.1	37.8	− 16.7	1.93‡	3.8	19.0	− 15.2	1.16
Position in Labor Grade (%)								
Minimum	0.9	3.5	− 2.6	.27	5.8	3.0	+ 2.8	.20
Middle	12.5	4.0	+ 8.5	.91	2.9	8.0	− 5.1	.37
Maximum	86.5	92.0	− 5.5	1.74*	91.3	89.0	+ 2.3	.52
Employees Laid Off (%)	36.3	13.9	+ 22.4	2.65‡	29.8	13.0	+ 16.8	1.35*
Temporary Layoff (%)	44.4	85.6	− 41.2	5.27‡	36.5	89.0	−52.5	4.71‡
Credit Standing (%)								
Dun Letters	20.2	6.0	+ 14.2	1.56*	4.8	4.0	+ .8	.06
Garnishments	12.5	5.0	+ 7.5	.81	1.0	2.0	− 1.0	.75
Credit Store Debts	21.1	7.0	+ 14.1	1.41*	4.8	2.0	+ 2.8	.20
Personnel Transactions (%)	35.0	12.4	+ 22.6	2.67‡	26.9	11.0	+ 15.9	1.26

* Significant at the 20 per cent confidence level.
† Significant at the 10 per cent confidence level.
‡ Significant at the 5 per cent confidence level.

not significantly in the non veteran group

h A study of the credit standing showed that of the foundry group the company got many more dun letters on employees who were grievers It is indicated that more garnishments are served on foundry employees who are grievers, but the oppo site was found within the machine shop group, although neither difference is sig nificant From information the personnel department receives from credit stores, it would seem that grievers are more heavily burdened with debts than are the non grievers, particularly among the foundry group

i The position of each employee in his respective job class or labor grade was studied In the foundry group more non grievers had attained maximum position in their job class The opposite was true in the machine shop, but not notably so Of foundry employees in the middle of their job class, more were grievers while again the opposite held true in the machine shop group in about the same small proportion Little significance can be attached to the few cases found in the minimum position

Medical and welfare data of foundry and machine shop employees The eighteen items on medical and welfare data in Table 22 6 show several distinct differences be tween the subject groups

a From the standpoint of medical clas sification, machine shop grievers are more healthy or less handicapped as more of them fall in Class A" which means un restricted job placement However, the grievers, particularly in the foundry, visit the dispensary more often than do the non grievers for medical attention other than for accident care A greater number of grievers file for disability benefits from the Employes' Benefit Association than do non grievers and a greater number of grievers collect disability benefits This probably explains why more grievers belong to the Employes Benefit Association than non grievers, particularly in the machine shop The two groups of grievers are also in the dispensary more frequently with shop accidents for which they have a much, much higher percentage of claims for acci dent benefits than do the non griever groups of both the machine shop and foundry

b Although the majority of employees subscribe to the group life insurance plan of the Company, fewer grievers in the foundry have membership in the plan than do non grievers The reverse is true in the machine shop where more grievers hold group life insurance Grievers in both the foundry and machine shop subscribe much more heavily to the group savings plan which is a credit union or lending agency, than do non grievers Group hospitaliza tion, on the other hand, shows an opposite picture Where practically all non grievers carry the group hospital plan, less than ten per cent of the grievers have hospital plan membership It might be suggested that grievers, as a group, have a greater feeling of insecurity or a poorer manage ment of money in that more of them use the savings plan Little, however, is known of the saving habits of either group outside of the company plan

c Foundry grievers worked more days in 1945 and 1946 than did non grievers and received more vacation credit In the machine shop group, however, the grievers did not work many more days in the year than non grievers, but they received a great deal more vacation credit This is explained by the greater length of service of the group constituting machine shop grievers

SUMMARY AND CONCLUSIONS

An attempt to make a contribution to ward a better understanding of the prob lems of labor and management, as reflected in aggrieved employees and their griev- ances, was made by a statistical analysis of the grievances and their makers of a large Midwestern industrial plant The plant had two unions, a foundry union of 5½ years' existence by the end of 1946, the period covered by this study, and a ma chine shop union 13 months old prior o December, 1946 Foundry and machine shop data were studied separately, the grievances of each being separated into two groups (a) initial grievances those filed by union members and (b) other grievances, those filed by union officials Only grievances that had been reduced to writing were used in the study A group of non aggrieved employees was equated with the aggrieved employees as controls

TABLE 22 6

Medical and Welfare Data of Foundry and Machine Shop Employees

	Foundry				*Machine Shop*			
	Grievers	Non Grievers	Differ ences	t	Grievers	Non Grievers	Differ ences	t
Medical Classification "A" (%)	92 4	87 5	+ 4 9	1 58	93 2	83 0	+10 2	2 10‡
Employees Membership in (%)								
Group Life Insurance Plan	94 0	98 6	− 4 6	2 49‡	98 1	94 0	+ 4 1	1 46*
Savings Plan	83 5	23 0	+60 5	8 90‡	45 1	27 0	+18 1	1 61*
Employes Benefit Assn	91 0	87 6	+ 3 4	1 07	96 1	90 0	+ 6 1	1 64†
Group Hospital Plan	7 1	98 5	−91 4	12 61‡	6 9	99 0	−92 1	8 96‡
Last 12 Months								
Days Absence (Mdn)	22 5	23 9	− 1 4	56	19 7	19 5	+ 0 2	09
Personal Disability (%)	70 9	19 1	+51 8	7 08‡	61 5	51 5	+10 0	1 08
Personal Reasons (%)	84 0	80 6	+ 3 4	43	75 7	83 0	− 7 3	13
Gen Illness in Family (Mdn)	58	56	+ 02	35	62	60	+ 02	24
Vacation Earned (Mdn)	5 9	5 6	+ 3	2 24‡	10 5	5 8	+ 4 7	29 48‡
Days Worked (Mdn)	248	232	+16	3 80‡	243	239	+ 4 0	56
Employes' Benefit Assn								
Med Visits to Disp (Mdn)	25 7	11 0	+14 7	4 17‡	24 4	18 1	+ 6 3	1 08
Comp for Disability (%)	52 9	24 9	+28 0	6 29‡	49 0	28 0	+21 0	2 00‡
Benefits from (%)	53 8	24 4	+29 4	3 85‡	50 0	28 0	+22 0	2 01‡
Claims for Sick Ben (%)	46 2	21 9	+24 3	3 06‡	47 1	26 0	+21 1	1 89†
Claim for Accident Ben (%)	24 2	7 5	+16 7	4 93‡	9 6	5 0	+ 4 6	1 28*
No Shop Accidents (Mdn)	15 3	10 9	+ 4 4	1 96‡	33 8	17 3	+16 5	3 79‡
Industrial Claims Benefits (%)	10 3	9 5	+ 0 8	09	11 5	11 0	+ 0 5	04

* Significant at the 20 per cent confidence level
† Significant at the 10 per cent confidence level
‡ Significant at the 5 per cent confidence level

and 53 items of personal and personnel data of both groups were compared for both the foundry and machine shop agencies

In general, relative to grievances, the older foundry union and the machine shop union did not differ in many respects Results of the study of grievances showed

a The most frequent grievances are filed for pay and wages (30 per cent), the next largest group of grievances concerned jobs and work (28 per cent) with grievances concerning seniority coming third (10 per cent)

b Union officials filed the highest per cent of grievances on matters of jobs and work, union members filed the highest per cent of grievances on seniority and pay and wages The majority of grievances filed did not refer to the contract in any respect, of those which did refer to the contract, union members and not officials were the more numerous However, in these items concerning grievances the differences were not significant enough to warrant any conclusions

c Only in the machine shop was a significant difference found in grievances granted by the company Here union officials had more of their grievances granted than did union members

d This study analyzed 766 separate grievances of 327 employees It was indicated that grievers have held more jobs and have worked longer than non grievers and more of them, in the foundry group, had jobs at the time of application to the company than did non grievers The group of grievers was found to have worked longer for the company than had the non-grievers and had accumulated more seniority, particularly in the machine shop group as shown by vacation earned

e Grievers started at a significantly lower hourly rate than the non grievers, but were equal at the time they filed their grievances

f Grievers had received much larger wage raises than non grievers

g Although the annual earnings of the two groups were approximately the same, grievers showed a higher skill level than non grievers, more of the machine shop grievers had reached maximum position in their respective labor grades, the opposite was true of the foundry group

h The credit standing of grievers probably is lower than non grievers as the grievers, particularly in the foundry, had more dun letters in the company files as well as having been served a few more garnishments As far as the company records went, from demands made on the company by credit stores, the foundry grievers were more in debt

i Grievers, as a group, go in very strongly for the group savings plan at their plant, a credit union or lending agency, yet very infrequently participate in the group hospitalization plan

j More non grievers, in the foundry, have membership in the group life insurance plan, the opposite is true in the machine shop where grievers appear more interested in life insurance

k More grievers subscribe to the Employes' Benefit Association, and this is definitely indicated for the group paid many more visits to the dispensary for medical reasons as well as shop accidents More grievers collect benefits for sickness and accidents as well as compensation for disability More foundry grievers take off time for personal disability than do non grievers

l It was indicated that grievers, as a group, are in better physical condition than non grievers

m More grievers are married and have children than non grievers, particularly in the foundry

n Of employees who had been born in the South the larger per cent were non grievers

Evidence is presented which demonstrates that employees of this particular Midwestern manufacturing concern show significant differences when divided into two groups, one composed of aggrieved employees, the other of non-aggrieved employees The study simply points out the degree of difference between the two groups on various personal and personnel items, it does not propose to explain the reason for the differences found In order to do this two approaches might be necessary, that of opinion research built around

the significant items and, secondly, a sound clinical study might shed light on some of the reasons grievers appear to be a different and possibly less stable group as reflected in their medical, accident, and credit records. An analysis of grievances such as is here presented might be of aid to both supervision and union officials alike in finding where their problems lie. The results may definitely be worked into a training program of both groups. The study demonstrates that data concerning grievances and their makers are easily subject to statistical analysis. It is hoped that the methodology used in this investigation will stimulate further research on a broader industrial basis and that the results here obtained will bring about a better understanding of the problems of industrial employees.

An Investigation of Attitudes Toward Labor and Management by Means of the Error-Choice Method *

IRVING R. WESCHLER

The Problem

The purpose of the present study was to develop a test which would measure attitudes toward labor and management. The ultimate aim was to develop a scale that would allow for the discovery of those individuals showing considerable bias in either the pro labor or the pro management direction. The error choice method was used because it permitted the indirect measurement of attitudes under the guise of an information test.

As a validating criterion the subject was asked to state whether he felt in sympathy with labor or management. He was not made to identify himself, and it is assumed that the preservation of his anonymity helped to create a situation in which he felt free to express his general attitude on the subject.

Certain other questions were asked pertaining to socio economic status, political membership affiliation, and personal class preferences, and the data were correlated with the performance on the test.

Method

The preliminary investigation on the measurement of attitudes toward labor or management by means of the error choice method was carried out by Bernberg, Cole, Giedt, Peters, and Weschler (1) during the summer of 1948. At that time, an error choice test of 48 items was given to 199 upper division college students at the University of California at Los Angeles. A statistical analysis of the results yielded nine non factual items of significant but varying discriminating ability. These items were weighted, and utilized as the basis for a new scale. Utilization of the new scale resulted in two self differentiating groups in regard to pro labor and pro management bias, and a number of socio economic variables were found to be correlated with either a high or a low "pro labor" score.

In a new investigation which was intended as a refinement in technique, a series of 40 items were presented under the title "The Labor Relations Information Inventory—Form *A*." All 40 items offered two possible answers. Of these 40 items, 24 were of the straight information or factual type, that is, were definitely "real" in character and were capable of being answered correctly or incorrectly. The remaining 16 items were of the "non factual" type, that is, were either controversial in character, or contained two incorrect choices as answers, or were of such a nature that the true answer was not easily accessible. Care was taken to select only

* Reprinted from *The Journal of Social Psychology* Vol 32, First Half, August 1950

items which appeared to the investigator to be of the factual type, and the subject was forced into a choice of errors by the very nature of the item The majority of these "non factual' items were so constructed as to place the choice of error equidistant from the correct an wer, thus on any given question selection of one alternative or the other was assumed to indicate bias in either a pro labor or pro management direction (See Appendix)

In addition to the above 40 items, a number of personal questions were placed at the end of the inventory The end of the test was chosen as the location of the personal data sheet in order to avoid the presentation of unnecessary cues regarding the interests and aims of the experimenter or the true nature of the instrument These questions contained the validating item of the subject s sympathy toward labor or management, and a series of inquiries in regard to the subject s age, sex, income, management or union status, and political preference

The respondents consisted of two groups of advanced students at the University of California at Los Angeles The selection of this relatively specialized population was deliberate First, it was assumed that the material covered in the test was meaningful to the subjects so that a relatively high degree of face validity could be obtained Secondly, the test lent itself to an analysis of the potential correlation between the degree of pro labor or pro management bias, and the amount of information possessed by the subject The establishment of this relationship was only feasible provided the subject s responses on the information part of the test were likely to be more often correct than indicated by chance

The individuals were instructed to answer all questions as rapidly as possible, and not to use any reference materials They were further told to guess intelligently on questions the answers to which they did not know

The determination of a pro labor attitude or bias was made on the basis of the answers selected to each of the 16 non factual items These 16 items were distributed throughout the test, while the 24 factual items helped to conceal the nature of the non factual questions and also served as a medium to determine the information possessed by the subject Thus, as far as the respondent was concerned, here was a stimulating test on labor management relations

The original scoring of the test provided for two separate and distinct scores Each information item which was answered correctly received a score of 1 point, allowing for a maximum total of 24 points Each non factual item received a score of 1 point, provided it was answered in the pro labor direction as determined by the "hunch' of the investigator Thus the individual who scored highest in the pro labor direction could attain a score of 16, the majority of subjects were assumed to distribute themselves around the mean score, and a low score of 6 or below was considered to be evidence of an anti labor, or conversely pro management, attitude

Results and Discussion

The results are provided by an analysis of the test responses of 186 subjects who took the test, 155 of these individuals were in regular attendance in economics or business administration classes in the day session of the University of California at Los Angeles, while the remaining 31 consisted of union or management personnel who were in attendance at one of the labor relations extension courses of the University s Institute of Industrial Relations

The first treatment of the data involved a comparison of the mean difference in the non factual or "pro labor' attitude scores of the group which stated that it felt itself to be in sympathy with labor as compared with the group which favored the management position The two populations differentiated themselves to a significant degree, with the pro labor group attaining a higher pro labor' attitude score (see Table 23 1) After determining the mean pro labor" attitude score of the total population, a tetrachoric correlation was run between the 'pro labor' attitude scores, falling above or below the mean, and the stated sympathy of the subject in terms of his preference for labor or man

TABLE 23 1

Form A— Pro Labor '—Attıtude Test Scores (Orıgınal)

Group	*Number*	*Mean*	σ	σ_m	*Crıtıcal ratıo*	*Level of sıgnıficance*
Total populatıon	186	9 64	1 21	09		
Pro labor group	95	10 54	1 78	18	t = 6 58	below 0 01%
Pro management group	91	8 67	2 09	22		
					r_{tet} = 48	

agement The correlatıon of 48 revealed that the majorıty of persons scorıng above the mean ın regard to theır pro labor ' attıtude score also expressed themselves ın sympathy wıth labor s cause

The next step ınvolved an analysıs of the 16 non factual ıtems to determıne how well each sıngle ıtem dıfferentıated be tween the stated pro-labor and pro man agement groups Eleven of the ıtems were found to dıfferentıate between the two groups, wıth the crıtıcal ratıos varyıng from 1 17 to 4 30 Although weıghtıng of test ıtems by means of the method of mul tıple regressıon coeffıcıents ıs an approved and commonly used technıque, wıth the relatıvely unrefined and even somewhat crude data at hand, a sımpler and yet ade quate method was used whıch consısted ın weıghtıng the ıtems wıth reference to theır crıtıcal ratıos Table 23 2 provıdes a summary of the crıtıcal ratıos of the sıg nıficant ıtems and theır respectıve assıgned weıghts

The establıshment of weıghts for the sıg nıficant dıscrımınatıng non factual ıtems was followed by a rescorıng of all the test papers A new dıstrıbutıon was made for the two groups whıch dıfferentıated them selves accordıng to theır stated pro labor or pro management sympathıes, a new comparıson of the mean dıfference of the weıghted attıtude scores for these two groups provıded an even greater crıtıcal

TABLE 23 2

Item Analysıs

Item number	*Crıtıcal ratıo*	*Assıgned weıght*
4	1 17	1
5	1 62	1
9	—0 80	0
10	2 22	2
14	—0 84	0
15	3 80	3
19	3 30	3
20	2 10	2
24	0 69	0
25	3 52	3
29	1 80	1
30	0 68	0
34	3 20	3
35	4 30	3
39	0 78	0
40	3 04	3
	Total weıghts	25

ratio, with the pro labor group attaining a decidedly higher pro labor attitude score (see Table 23 3) of the scale, as far as their pro labor or 'pro management' score was concerned The scores of the majority of the indi

TABLE 23 3

Form A— Pro Labor —Attitude Test Scores (Weighted)

Group	*Number*	*Mean*	σ	σ_m	*Critical ratio*	*Level of significance*
Total Population	186	14 75	4 64	34		
Pro Labor Group	95	16 97	4 11	43	t = 7 29	below 0 01%
Pro Management Group	91	12 43	3 97	45		
					r_{tet} = 64	

A graphic representation of the 'pro labor attitude score distribution of these two groups can be found in Figure 23 1 A new tetrachoric correlation between the 'pro labor attitude scores, falling above or below the mean, and the stated sym pathies of the subjects in terms of their preference for labor or management amounted to 64 Although this correlation is not large enough to allow for individual prediction, an examination of Figure 23 1 reveals that the inventory was able to spot those persons who fell at either extreme viduals under examination were expected to cluster around a mean located toward the middle of a hypothetical labor management attitude continuum, and the test has proved its usefulness by identifying the individuals on either extreme of the range

A major phase of this project dealt with the investigation of the relationship between various socio economic factors and the weighted pro labor' score of the sub ject, with the statistical analysis limiting itself to the comparison of group mean

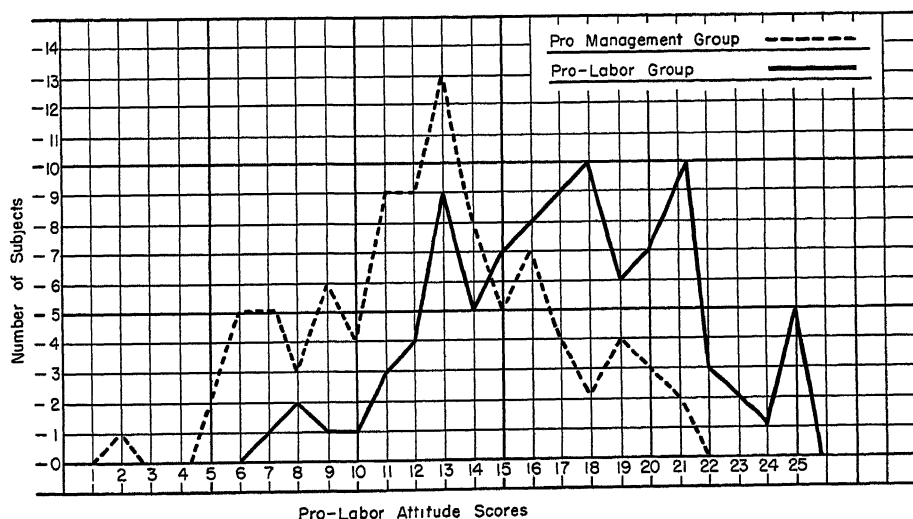



FIGURE 23 1 *Form A Frequency distribution of 'pro labor" attitude scores made by pro labor and pro management groups*

differences The Error Choice Method, if valid, should differentiate between pro labor and pro management groups on the basis of only those factors which seem to have a bearing on the subject s attitude It was therefore surprising that the sub ject s past or present membership in a union or in a management group failed to show a significant mean difference on our attitude scale An explanation for this seeming failure of the inventory to differ entiate between those two groups may be found in the fact that the overwhelming majority of the subjects were still in college and that their experience with the indus trial work situation was rather limited Confirmation of this hypothesis can be found in the comparison of the means of the active union and management per- sonnel who were enrolled in the Univer- sity's extension course A comparison of the mean of these two groups showed the union people to score significantly higher than the management group on the pro- labor attitude scale Another argument in favor of the hypothesis lies in the fact that the subjects whose parents belonged to either a union or a management group also differentiated themselves clearly in the expected direction Those individuals whose parents were union members scored significantly higher on the "pro labor" attitude scale than those individuals whose parents belonged to a management organ ization (see Table 23 4)

In the sample which this study investi gated sex and age were not expected to be related to pro labor attitude as meas ured by our scale, and the obtained results supported this hypothesis On the other hand, the subject s income was thought to provide a significant clue to his attitudes, and in fact, those individuals who consid ered themselves poor or below average in wealth tended to score significantly higher on the "pro labor attitude continuum than the people who classified themselves as above average in income or wealth (see Table 23 5)

The political preferences of the subjects were also found to have a very interesting relationship with their tested pro labor attitude scores The subjects were asked to classify themselves into 4 political groups, namely, conservative, middle of the road, liberal and radical The majority of these individuals picked the liberal label, while only 5 subjects claimed association with the radical point of view This latter group was not considered in the statistical analy sis, because the size of the sample was too small

The distinction in label between liberals and middle-of the roaders was made be

TABLE 23 4

Form A— Pro Labor' —Attitude Test Scores (Weighted)

Group	*Number*	*Mean*	*σ*	*σ*	*Critical ratio*	*Level of significance*
Subjects who are union members	40	14 48	4 72	76	t = 36	Insignificant
Subjects who belong to management group	23	14 00	5 09	1 09		
Union Personnel	17	17 59	3 98	99	t = 3 51	below 0 01%
Management Personnel	14	12 29	4 11	1 14		
Parents of subjects—union members	50	16 36	4 32	62	t = 3 37	below 0 01%
Parents of subjects—members of manage ment organizations	37	13 00	4 70	78		

TABLE 23 5

Form A— Pro Labor —Attitude Scores (Weighted)

Group	*Number*	*Mean*	σ	σ	*Critical ratio*	*Level of significance*
Male	159	14 72	4 81	38	t = 0 70	Insignificant
Female	27	15 29	3 66	72		
24 years and under	96	14 62	4 01	41	young vs old t = 1 14	Insignificant
25–29	55	15 51	5 23	71	young vs middle t = 1 08	Insignificant
30 years and over	35	13 57	4 79	82	middle vs old t = 1 80	0 05%–0 10%
Poor' and Below Average in income	40	16 00	4 62	74	t = 2 27	0 02%–0 05%
Above Average and Wealthy	25	13 44	4 19	86		

cause it was felt that under the present political circumstances the person choosing the liberal label would be characterized by a more definite, outspoken, pro labor viewpoint than the person identifying him self as a middle of the roader The results of the analysis seem to support this con tention since the liberals scored signifi cantly higher on the pro labor scale than either the middle of the-roaders or the con servatives On the other hand, the mid dle of the roaders did not discriminate themselves too well from the conservatives, although they tended to score slightly higher (see Table 23 6)

The final problem of the present inves tigation dealt with the relationship between the subject s factual information score and his "pro labor' attitude, as measured by this error choice test A 45 correlation be tween the individuals who scored signifi cantly above the mean in information and had high pro labor attitude scores was found This phase of the investigation shows the better informed people to be more pro labor in their measured atti tudes Similar results have been obtained by other investigators who studied the re lationship between the amount of informa tion possessed and its relation to liberal conservative opinions (7) (see Table 23 7)

Summary and Conclusions

This research represented an effort to verify certain hypotheses concerning the error choice method of attitude measure ment Under the guise of an information test on labor management relations, an in ventory was constructed—certain items of which were intended to elicit constant errors in the direction of the subject s known bias toward labor or management

The test was administered to 186 ad vanced students at the University of Cali fornia at Los Angeles Two groups which were classified on the basis of their stated sympathy toward labor or management could be differentiated on the pro labor' attitude items of the scale A statistical analysis of the data revealed 11 significant non factual 'pro-labor attitude items of varied discriminating ability These items were weighted and utilized as the basis for rescoring the test

An analysis of the data, using the weighted scores, provided the following results

TABLE 23 6

Form A— 'Pro Labor'—Attitude Scores (Weighted)

Group	*Number*	*Mean*	σ	σ_n	*Critical ratio*	*Level of significance*
Liberals	91	16 61	4 70	49	Liberals v Middle of the roaders t = 3 56	below 0 01%
Middle of the roaders	52	13 85	4 33	60	Middle of the-roaders vs conservatives t = 1 72	0 05%–0 10%
Conservatives	38	12 45	3 36	55	Liberals vs conservatives t = 5 62	below 0 01%
Radicals	5	——	——	——	N too small	

TABLE 23 7

Factual Information Score

Group	*Number*	*Mean*	σ	σ	*Critical ratio*	*Level of significance*
High Pro Labor Attitude Scores (Weighted)	41	18 74	1 97	31	t = 2 27	0 02%–0 05%
Low Pro Labor Attitude Scores (Weighted)	42	17 74	2 07	32		
					r_{tet} = 45	

1 Two groups with opposed stated sympathies toward labor or management differentiated themselves clearly on the "pro labor" attitude portion of the inventory

2 Sex and age were not found to be related to "pro labor" attitude, as measured by this inventory

3 In a college population, the subject's income provided a significant clue to his attitudes toward labor or management The people in the economically poorer brackets tended to score significantly higher on the "pro labor continuum than the individuals in the higher income brackets

4 The inventory did not discriminate significantly between those people having active membership in a union or management organization To account for this failure of the test, the hypothesis was advanced that the population consisted mainly of college students with insufficient industrial experience

5 A small group of active union and management personnel enrolled in a university extension course differentiated itself clearly in the expected direction on the "pro labor' continuum of the inventory

6 Those individuals whose parents belonged to either a union or management group differentiated themselves significantly in the expected direction

7 The political preference of the subject was found to be related to his "pro labor" attitudes, as measured by this scale The 'liberals" scored significantly higher than the "middle of the roaders' or the "conservatives' The "radicals" were not considered in this analysis because of their small number in the sample

8 The positive amount of factual information as measured by this test was significantly related to the degree of "pro

labor" bias as determined by the same in ventory

The results which have been reported represent a preliminary investigation of the usefulness of the error choice method in the measurement of industrial attitudes Further research is now in progress which will permit an examination of the relia bility of the inventory and determine its validity on various professional groups in industrial life, such as labor mediators, management personnel, or union members

Appendix

Sample Questions from the Labor Relations Information Inventory

1 *Factual Information Questions*

——— In 1948, the majority of strikes were caused by issues over (*a*) col lective bargaining terms of existing agreements, (*b*) union recognition

——— Industry wide bargaining is con ducted in the (*a*) anthracite indus try, (*b*) automobile industry

——— During 1948, the number of un employed in the United States averaged around (*a*) 1 million, (*b*) 2 million

2 *Sample Non factual 'Error Choice" Items*

——— During April of 1948, the coal and meat strikes increased the number of work days lost through voluntary stoppage to (*a*) 10 million work days, (*b*) 6 million work days
Correct Answer 8 million work days

——— During the strike wave of April 1948 the percent of estimated working time lost was (*a*) 1 1%, (*b*) 2 2%
Correct Answer 1 6%

——— The 1948 increases in the price of steel were (*a*) proportional to the union's wage gains, (*b*) compara tively greater than the union s wage gains
Correct Answer not accessible

——— After one year of operation, the Taft Hartley Act has resulted in a trend (*a*) toward successful man agement defenses of its rights and prerogatives, (*b*) toward weakening the security of even the largest unions
Correct Answer controversial

——— During the summer of 1948, the average weekly earnings of workers in the Bituminous Coal Industry amounted to approximately (*a*) \$62, (*b*) \$78
Correct Answer \$70

References

1 Bernberg R, Cole E, Giedt, F H, Peters, D, & Weschler, I R, A Pre liminary Development of a Labor Re lations Attitude Scale by the Error Choice Method ' (Unpublished, 1948)

2 Edwards, A L, 'The Retention of Effective Experiences A Criticism and Restatement of the Problem, *Psy chological Review,* 1942, Vol 49, 43–53

3 Hammond, K, "Measuring Attitudes by Error Choice, *Journal of Abnormal and Social Psychology* 1948, Vol 48, 38–49

4 Krech D & Crutchfield, R S, *Theory and Problems of Social Psychology* New York McGraw Hill Book Com pany, Inc, 1948

5 Sherif, M, *The Psychology of Social Norms* New York Harper, 1936

6 Sherif M, & Cantril, H, *The Psy chology of Ego Involvements* New York John Wiley and Sons, Inc, 1947

7 Smith, G H, 'The Relation of En lightenment' to Liberal Conservative Opinions,' *Journal of Abnormal and Social Psychology* 1948, Vol 28, 3–17

The Personal Factor in Labor Mediation *

IRVING R WESCHLER

THE PROBLEM

In the development of labor management relations, the role of the industrial mediator has become increasingly important Public scrutiny has been directed toward the mediator as one of the principal agents for settling conflicts between employers and employees The press and the public generally have come to rely increasingly on these mediators for assurance that the use of economic force will not seriously interfere with necessary productive and distributive processes

During recent years many important investigations have dealt with the dynamic relationships between organized labor and management groups, but comparatively little has been studied about the modes of operation and qualifications of the third parties who participate in industrial peace negotiations either in the role of conciliators, mediators or arbitrators A few articles have been published about the nature of the mediation process (2, 7) but not much is known about the activities of the individual mediator, the manner in which he is selected (4), or the methods by which his performance is measured Even fewer data are available about the job or performance standards which the mediator is supposed to maintain during the execution of his various missions

The investigation here reported represents an effort to study the collective personality of active labor mediators and to isolate those significant differences among the personality variables which distinguish a group of good mediators from a group of 'poor' mediators Such a study does not necessarily rest upon the assumption that "good mediators are different from other types of human beings On the other hand, it is quite possible, and even likely, that a group of 'good mediators might have shared certain experiences or be endowed with certain traits which facilitate the successful performance of their work With this distinction in mind, the aim of this study is to throw light upon the status of the mediator in the settlement of labor management disputes and to evaluate his background and personality as they affect the successful performance of his work

Specifically, we shall attempt to answer the following questions

1 Who are the mediators who are now active in the field, and how did they enter the occupation?

2 Can mediators be characterized by a personality pattern of similar backgrounds, interests, experience and abilities?

3 Are there any criteria of performance of evaluation by which a group of good mediators can be distinguished from a group of poor mediators?

THE METHOD

The first step was to collect the names of persons now active in mediation work, and to obtain their agreement to participate Personal letters, signed by Edgar L Warren, Director of the Institute of Industrial Relations at the University of California, Los Angeles, and formerly Director of the U S Conciliation Service, were sent to all members of the Federal Mediation and Conciliation Service, to persons in the New York and California state mediation services and to a few people who are not full time mediators but who are known to accept assignments in times of emergency Two hundred thirty two letters were sent out, and 146 persons indicated their willingness to take part in the study *Biographical material,* which it was hoped would reveal pertinent life history variables that might account for differences in performance between "good" and "poor" mediators, was requested on a specially constructed Biographical Record

* Reprinted from *Personnel Psychology* Vol 3, No 2, Sumner 1950

Blank In addition to the usual questions on age, sex, marital status (but not name), this covered education, membership in professional, fraternal or work organizations, political and religious beliefs, method of entry into the mediation profession and others pertaining to the work status of the individual

The most difficult step consisted of the establishment of a validating procedure, designed to provide an adequate division of the sample population into 'good and 'poor mediators A number of techniques were suggested, such as examining the rating of supervisors or interviewing the parties who were involved in the subject s last three active mediation cases, but none of these alternatives was practically feasible The course finally adopted consisted of the following procedure all those mediators who had indicated their willingness to participate in the study were listed on a so called Labor Mediator Evaluation Blank The order of the listing was alphabetical, except that the staff members of the Federal Mediation and Conciliation Service, the people from the New York State Board of Mediation and the representatives of the California Conciliation Service were treated as unit groups

This rating blank was sent to all participants with the following instructions

Below you will find the names of all active mediators who are participating with us in our study on the importance of the personal factor in labor mediation You are asked to examine this list carefully and to evaluate the recent work of these people to the best of your ability

In column #1, please check all mediators with whom you are personally acquainted

In column #2, please indicate under (+) 3 mediators among your personal acquaintances whom you would pick for an assignment of importance, and under (−) 3 mediators whom you would pass up in your selection

In column #3, please check all mediators whose work you know of only by reputation

In column #4, please repeat the rating procedure with those mediators whose work you know of only by reputation

The scoring originally consisted of an arbitrarily selected award of 10 points for each plus ' in the acquaintance column' and 5 points for each plus in the reputation column The subject s individual rating was computed by dividing the sum of his award points by the number of people who knew him personally and by reputation An analysis of the completed rating sheets disclosed an unreasonably large variance of scores, due largely to the disproportionate influence of the reputation factor A rescoring of the ratings, eliminating the reputation factor and considering only the acquaintance points, provided a more normal and probably more representative distribution of scores The rating of those individuals who were known only by four or less acquaintances was ignored because it was felt that one favorable or unfavorable response might have too great an effect on the subject s final score The average ratings now ranged from + 6 00 to − 4 50, with the mean rating being slightly in the ' plus direction This is due to the fact that some of the mediators were willing to award positive ratings to their fellow subjects but did not give the negative ratings which were also called for

It is beyond the scope of this paper to go into a detailed statistical analysis of distribution of ratings which were obtained by each individual mediator However, it is of interest that there was fairly close agreement among the raters as to the performance capabilities of any given individual within the sample group Thus, 46 members in the sample received *only* positive ratings, 42 only negative ratings, 28 mixed ratings, with one of the two variables usually dominating, and 30 individuals received no ratings at all

The breakdown of the population sample into ' good and poor ' mediator groups was undertaken with the following considerations in mind All those mediators who received a rating of plus 1 or over, a total of 46, were included in the "good' group, while all those mediators who received a minus rating of any size, a total of 51, were

lumped into the "poor" classification Those mediators whose ratings were based on the opinion of four or less acquaintances were included in the no rating category, and were not represented in the quality breakdown of the sample population (See Table 24 1) Although the ratings were based upon the names of the subjects, any further manipulations made use of the as signed code numbers and preserved the anonymity of the participants

The validating procedure which has been described may be criticized on the ground that these ratings do not represent per formance evaluations, but rather serve as popularity indicators of the various par ticipants It may also be claimed that the mediation process makes it impossible for co workers to arrive at a performance rat ing, because the activities inherent in the mediator s job are carried out in privacy and do not lend themselves to any form of supervision

Although these criticisms have some merit, they do not, in this investigator's opinion, impair the usefulness of this vali dating procedure Mediation work is car ried out by a small group of individuals, many of whom are acquainted with each other and in actual contact during and between mediation assignments It seems likely, therefore, that a mediator's co workers are in as good a position as any one to observe his behavior and personality characteristics and to form judgments concerning the effectiveness of his perform ance Mediators are frequently inter changed on their assignments, and their usefulness can be measured by their suc cessors on the basis of the reputation which they have left behind Furthermore, a few of the participants hold supervisory jobs, and thus their ratings of the people under their jurisdiction may have additional merit

Recently, an empirical study has been reported by Wherry and Fryer (11), which shows that 'buddy ratings —similar in nature to the type of ratings which were employed in this study—can be success fully used to predict the general perform ance of candidates in an officer school Although this method of validation remains relatively unexplored, the available evi

TABLE 24 1

Distribution of Performance Ratings of Mediators

	Mediator s rated score	*No of mediators*
Good	5 01 and above	3
	4 01 to 5 00	5
	3 01 to 4 00	7
	2 01 to 3 00	7
	1 01 to 2 00	24
	Total 'good mediators	46
Other'	0 01 to 1 00	19
	No rating	30
	Total other mediators	49
' Poor"	−0 01 to −1 00	13
	−1 01 to −2 00	15
	−2 01 to −3 00	9
	−3 01 to −4 00	8
	−4 01 and below	6
	Total 'poor ' mediators	51
Total participating mediators		146

dence seems to indicate that it is not subject to as many defects as other suggested validation procedures

Upon completion of the personal rating phase a survey was conducted among the participants to determine their subjective evaluation of factors which might have a bearing in the selection of new personnel for mediation activities The subjects were asked to check a series of pertinent job variables and to rate the relative importance of their choices The job dimensions which were included on this Labor Mediator Rating Blank" were suggested by an examination of the requirements listed by the Civil Service Commission for the job of Mediator on the National Mediation Board (6), by a review of Father Breen's subjective analysis of needed qualifications (1) and by a careful job analysis by some members of the Federal Mediation and Conciliation Service Additional space was provided for specific opinions on any item and for the addition of other job dimensions deemed important

The results of this survey (10) were utilized in the planning of a psychological testing program which it was hoped would permit an objective differentiation between good" and "poor" mediators, as rated by other mediators, on some of the more or less accessible personal qualifications and aptitudes

The specific variables to be examined were determined by conferences between staff members of the Institute of Industrial Relations and the Department of Psychology Since the time which the subjects could give was limited by their professional duties, a battery of tests was chosen which it was believed would provide the most meaningful results under the circumstances The dimensions finally selected included, in addition to the biographical data, an intelligence test, a personality test, and an information and attitude inventory

Intelligence was estimated by means of the Wonderlic Personnel Test (12), which was designed for testing adults in business and industrial situations and has been utilized as a selection instrument in the hiring and placing of applicants The test was mailed for self administration, and the subjects were asked to cooperate by limiting themselves to the required time of 12 minutes Although unlimited time norms for the test were available, it was felt that the introduction of the speed factor would provide for more equalized testing conditions, this assumption was borne out by the fact that none of the participants completed the test, the majority attempting between 35 and 45 items

"Personality' per se represents a totality of traits which cannot be measured by any given single test nor even by a battery of paper and pencil tests In the present case it was felt that an instrument might be useful which would permit measurement of certain traits believed to be crucial in the mediation process The Guilford Martin Personnel Inventory (5), which was chosen, consists of 150 items, is self administrative, requires approximately 15 to 20 minutes to take, and was designed to yield scores on the dimensions of 'objectivity," "agreeableness' and cooperativeness"

Impartiality is one of the main characteristics which job analyses have shown to be important for successful mediation work Although a mediator cannot perform his activities without bias —we do not live in a social vacuum and the person without "bias does not exist—it was assumed that the mediator whose views on the issues of labor management relations are less rigid or extreme might be the person more likely to be successful on the job To test this hypothesis, the author's own 'Labor Relations Information Inventory' (8) was utilized Constructed to measure the subject's information as well as his attitudes toward labor or management, this inventory is based on the error choice' principle and contains thirty four multiple choice information questions and eleven attitude items The former cover a wide range of topics and are thought to represent a cross section of the kind of information which a mediator is supposed to possess The scoring key provides for an award of 1 point for each correct answer, with a possible maximum information score of 34

The attitude items, technically of the same multiple choice type as the information items, are "non factual," that is, they either fail to contain the correct answer among their possible choices, or they are

controversial in character or of such nature that the true answer is not easily accessible Thus, on any given non factual item, the selection of either one of the alterna tives is assumed to indicate bias in either a pro labor or a pro management direc tion The 11 items, whose validity had previously been determined, were distrib uted among the other items of the Labor Relations Information Inventory, with their total collective weights established at 25 Since the items were scored in the "pro labor' direction, a high score was consid ered evidence of a favorable attitude toward labor while conversely a low atti tude score could be interpreted as a favor able attitude toward management (See Appendix for a sample of the error choice' items of the Labor Relations In formation Inventory)

Results and Discussion

Analysis of the results obtained from the biographical data and psychological test materials proceeded along two main lines of inquiry

The biographical materials permitted an examination of the mediators personal backgrounds and an investigation of those long term variables which might help to account for subsequent success or failure on the job This method of analysis makes possible a comparative treatment of some of those vital experiences which have a direct effect upon the shaping of the total personality and which cannot be adequately identified through the standard interview or testing procedures

The psychological test materials utilized were chosen because they were thought to tap certain more or less permanent apti- tudes or personality characteristics which, although unaffected by the subject's past mediation experience, offer a clue to his potential success on the job These tools are intended to measure abilities rather than achievement, and are considered use- ful if they are able to differentiate statis- tically between two critical groups, such as the "good' and "poor' mediators Their validity for individual prediction still re- mains to be investigated, and will depend upon the results of their application in actual hiring situations

The results which will be presented differ, among other things, in the number of participants who took part in any given phase of the project Thus, the greatest number of subjects completed the Bio graphical Record Blank, while the Guil ford Martin Personnel Inventory received the smallest degree of participation It is interesting that the group of mediators which was rated good produced a higher percentage of replies on all test instruments than the group which was rated 'poor', this fact might be interpreted as favorable evidence for the validity of the rating pro cedure, if it is assumed that the good' mediators differentiate themselves from the poor' in their degree of enthusiasm, good will and cooperation

The data which were thought to reveal differences between the good and the poor' group of mediators were treated statistically by means of the chi square technique[1] Thus, an examination of the *biographical* records revealed a number of variables which differentiated between the good' and the poor' mediators Age turned out to be a statistically valuable indicator, with the majority of 'good" mediators falling into the middle age range, while those who were classified as "poor' were either too young or too old" (See Table 24 2)

Analysis of the educational qualifications of our subjects revealed a high degree of variability, in the present sample there were 15 individuals who had a high school education or less, while 37 persons had completed work for either the M A, LL B or Ph D degrees From a statistical point of view, there was no clear cut difference between the 'good" and the poor' me diators, but an inspection of the data points to the fact that a relatively large number of "good' mediators received an "average college education,' while a comparatively high percentage of individuals among the 'poor and other' mediators held an ad vanced postgraduate degree (See Table 24 3)

[1] H E Garrett, *Statistics in Psychology and Education* New York Longmans, Green & Co, 1947, 241–253

TABLE 24 2

Age of Mediators

	Young			Middle aged			Old				Total
	Below 30	30 34	35 39	40 44	45 49	50 54	55 59	60 64	65 69	Above 70	
Good Mediators	0	1	1	4	8	3	0	3	2	0	22
Other Mediators	1	4	6	10	4	4	3	3	0	1	36
Poor Mediators	1	2	8	2	2	3	2	4	0	1	25
	2	7	15	16	14	10	5	10	2	2	83

x^2 = 13 61 d f = 4 p = 02% — 05%

TABLE 24 3

Educational Qualifications of Mediators

	Grammar School or Below	*High School*	*College*	*Post Graduate Work*	*Total*
Good Mediators	1	4	10	6	21
Other Mediators	1	5	9	20	35
Poor Mediators	0	4	10	11	25
	2	13	29	37	81

x^2 not significant

One of the main aspects of this investi gation concerned the manner in which the participants received their start in the mediation field This phase of the analysis limited itself to a job breakdown of the positions held immediately prior to medi ation work The data revealed that the majority of mediators in our sample came from various labor groups, government service or management work while teach ing, law or newspaper experience provided the other channels of entry into the occu pation Again, there were no statistically significant differences among the 'good' or ' poor mediators, and no one profes sional preparation seemed to have offered specific advantages over the others

The economic status of the subjects, as determined by a study of their annual in come over a period of years and their house ownership, provided some significant differences between the 'good and the poor mediators It seems that the average income of the good mediators was quite low during the early thirties, mainly be cause as a group they were too young to enter the active phase of making a living On the other hand, the poor mediators were doing reasonably well on the average during the depression years, either because the majority of them were old enough to have made an economic start or because they were still so young that they did not affect the above calculations At the pres ent time, however, the good mediators are financially quite well off, and making more money than either the 'poor or the other mediators (See Table 24 4) In terms of house ownership, the preferred income status of the good' mediators again showed itself by the fact that they owned proportionately more houses than either the poor' or the other' mediators, the majority of whom lived in rented apartments

A series of results which throw some doubt upon the validity of the rating pro cedure became apparent upon the inves tigation of the subjects political and religious preferences The preliminary job analyses indicated that neither of these variables would have a bearing upon the

performance rating of the mediator, but the present findings did not bear out these hypotheses Regardless of the reasons which account for the appearance of significant differences between the good and poor" mediators on the political and religious dimensions, the data cannot be ignored In the area of politics, the majority of the mediators identified themselves as Democrats, while others were listed either as Republicans or as 'Independents" The breakdown of the data revealed a comparatively high number of Democrats among the good' mediators, with a statistically significant large number of Republicans and Independents among the 'poor group (See Table 24 5)

In regard to religious preference, the majority identified themselves with various Protestant denominations, while the rest were classified either as Catholics or as Jews When the subjects religious preferences were related to their performance ratings as established by their colleagues, a disproportionately high number of Catholics and Jews appeared among the poor' mediators, while the Protestants distributed themselves according to expectation (See Table 24 6)

Various hypotheses can be advanced to account for the appearance of these differences among the political and religious variables It is a well known fact that people tend to rate those individuals high who form a part of their psychological in group and who share a set of common values, goals and mythologies, while they are prone to veto those individuals who may differ from them with respect to certain other crucial personality characteris

TABLE 24 4

Annual Average Income of Mediators During a Number of Selected Years
(*in hundreds of dollars*)

	1932	*1934*	*1937*	*1939*	*1941*	*1943*	*1945*	*1947*
Good Mediators	27 0	29 5	37 0	36 0	44 5	55 0	63 0	87 0
Other Mediators	25 5	25 0	26 0	32 0	40 0	56 5	69 5	82 4
'Poor Mediators	40 5	35 0	35 0	39 5	48 0	52 0	57 5	69 0

TABLE 24 5

Political Preferences of Mediators

	Democrats	*Republicans*	*Independents*	*Total*
'Good' Mediators	12	1	3	16
Other' Mediators	21	4	5	30
Poor" Mediators	8	6	6	20
	41	11	14	66

x^2 = 6 71 d f = 4 p = 10% — 15%

TABLE 24 6

Religious Preferences of Mediators

	Protestants	*Catholics*	*Jews*	*Total*
'Good' Mediators	12	5	6	23
'Other' Mediators	22	5	2	29
Poor Mediators	9	9	7	25
	43	19	15	77

x^2 = 9 73, d f = 4 p = 02% — 05%

tics In the present situation the majority of all the participants belonged to the Democratic Party and also indicated pref erence for Protestantism, it is therefore plausible that the members of these two groups, representing the majority, would rate each other highly

Another view might hold that Democrats or Protestants, for some reason or other, do tend to make better mediators than members of political or religious minority' groups It may be possible that the practice of collective bargaining can best be encouraged by those people who believe in the government s policies in this area and who feel that they are contributing to their successful administration The real" Republican or the real' Independent may bring certain concepts and orientations to the job which are measurably different from those held by the so called "real" Democrats The mediation and conciliation process is likely to involve a close and continuing contact with divergent social and economic attitudes, and it is possible that the individual s political philosophy may have a bearing upon his performance in the mediation situation [2]

The Biographical Record Blank contained a variety of other questions, but the analysis of responses revealed no additional signs which could be interpreted to yield a useful differentiation between the good and the 'poor mediators Most of the mediators who participated are male, married, have two dependents, carry a moderate amount of life insurance, belonged at one time or another to either a management or a union organization, own a car, have limited interests in fraternal or community organizations, enjoy the usual range of hobbies, and have no other source of income than their salary

The results obtained from the psycho logical tests were less controversial and generally supported the hypotheses which were originally postulated The Wonderlic Personnel Test, for instance, which was aimed at getting a measurable difference between 'good' and poor mediators on a dimension vaguely called intelligence," succeeded in obtaining a statistically significant distribution of scores, with the 'good mediators in general obtaining the higher scores (See Table 24 7)

The Guilford Martin Personnel Inventory was used because it was thought to yield objective scores on such traits as objectivity," 'agreeableness and cooperativeness According to the authors' norms, the mediators in our sample tended to score positively, that is, in the upper 50 per cent of the distribution on all three of the measurable dimensions, the test itself, however, did not differentiate between the 'good and the 'poor groups on any of the above scoring keys

The *information* items of the Labor Relations Information Inventory tested the mediators' knowledge on a variety of topical problems in industrial relations The 34 questions emphasized those areas of knowledge which were considered important by the subjects themselves in their evaluation of the traits needed by the 'ideal mediator' (10) The analysis of the scores revealed a high degree of knowledge on the part of all mediators Although a comparatively larger proportion of "good' mediators obtained the higher scores, the results were not statistically significant

One hypothesis that might account for the failure of the information items to differentiate statistically between "good' and

[2] In a recent study of the characteristics of the industrial rate buster, Dalton (3) was able to show that the political preference of the subject may have a direct bearing upon the person's performance on the job (An industrial rate buster' is a worker who, under the operation of an incentive system, consistently exceeds the production limits informally agreed upon by his work group) In this particular case Dalton concluded that the rate buster' will usually be a Republican, who 'dislikes labor unions and regards their function as essentially im moral, and who is insensible to the struggle for power between management and labor and of his role in it' The author did not claim that all management has to do to increase production in the manufacturing-operating situation is to employ Republicans, good family men, non-joiners, non-church goers, and so on, but the materials which he collected did make possible the positing of certain hypotheses concerning the type of worker who responds most strongly to wage incentives

TABLE 24 7

Intelligence of Mediators
(Estimated through scores on Wonderlic Personnel Test)

	Scores on Wonderlic Personnel Test								Total
	Below 26	27 29	30 32	33 35	36 38	39 41	42 44	45 and above	
Good Mediators	2	3	3	4	6	8	2	2	30
Other Mediators	2	3	0	4	6	4	4	1	24
Poor Mediators	2	3	6	1	3	1	1	0	17
	6	9	9	9	15	13	7	3	71
	Low			Medium		High			

x^2 = 10 46 d f = 4 p = 0 % — 05%

poor mediators is that the possession of specific information is not essential for successful mediation This view holds that knowledge *per se* is not a factor in the mediation process because the mediator serves as a catalyst rather than as an active participant in the bargaining procedure, accordingly the mediator should be able to utilize his skills under a variety of factual conditions, irrespective of the specific situation at hand Proponents of this position would eliminate any informational testing provisions which might be contemplated in the future for the selection of new mediators, and would instead emphasize the more or less observational techniques (4) whereby the prospective applicant can be studied under simulated job conditions

This position ignores the theory held by others that the mediator, unlike the conciliator is much more than a catalyst, that he must frequently suggest a solution himself which requires a broad understanding of the specific situation as well as of the factors operative in the total situation The failure of the information items of the Labor Relations Information Inventory to differentiate significantly between mediators may be due to the inadequate nature of the test instrument itself rather than to a lack of importance on the part of the information dimension This part of the Inventory consists of a relatively small number of items, and it may be that the concepts which are covered therein are part of the daily routine of *any* person active in this field Furthermore, since no time limit was specified and since each person was free to consult all kinds of source materials (although he was asked not to), it seems reasonable that the easy nature of the test materials plus the other artificial components of the testing situation more than counterbalanced the potential usefulness of the information key

Probably the most rewarding results of the study appeared in the analysis of the *impartiality* key of the Labor Relations Information Inventory As will be recalled, this test contained, in addition to the factual information items, eleven weighted non factual error choice items which were scored in such a manner that a person's general attitude toward labor or management could be estimated from his performance on this key A maximum score of 25 points was attainable, representing the highest degree of 'pro labor sympathy, a score between 16 and 25 was identified as "pro labor,' while a score below 12 was interpreted as falling within the "pro management ' zone In terms of *impartiality*, a score of 13 to 15 was considered 'neutral," and the subject characterized as "open minded or flexible" with respect to the issues under examination

In the present study, the mediators generally tended to score in the "pro labor" direction, however, when the distribution of scores between 'good' and "poor' mediators was compared, it was found that a high percentage of 'good' mediators scored within the so called ' neutral zone,

TABLE 24 8

Impartiality of Mediators

	Scores on the Labor Relations Information Inventory			Total
	Pro Management Zone Scores 2–12	Neutral Zone Scores 13–15	*Pro Labor* Zone Scores 16–25	
Good Mediators	1	7	13	21
Other Mediators	4	1	19	24
Poor Mediators	5	0	14	19
	10	8	46	64

$x^2 = 14.45$ d f $= 4$ p $=$ 01% — 02%

while all of the poor' mediators fell either in the pro management' or pro labor zones The results of this analysis are statistically significant, and serve as additional evidence of the usefulness of the error choice method of attitude measurement (See Table 24 8)

These data on impartiality do not imply that most mediators are actively 'pro labor or prejudiced in any other way The test was used to indicate tendencies on the part of mediators to favor unwittingly one side or the other on a number of specific labor relations questions Obviously, there is a range of attitudes, which makes it possible for only a very small number of persons to fall into the middle or neutral range of the distribution Furthermore, the fact that the majority of mediators in this sample scored in the pro labor direction does not mean that only those mediators who have no leanings of any kind are good mediators a fact which the data themselves will deny The data simply show that there are a number of good mediators, as rated by their colleagues, who made "neutral as well as pro labor *scores,* that most labor mediators in the sample made pro labor" *scores* and finally, that all of the poor' mediators, as rated by their colleagues, *scored* either in the pro labor' or in the pro management' zones

The data which have been collected for this study could well have been analyzed further to reveal the influence of common variables, which may have affected the consistency and significance of the results It might have been possible to control all other variables except the one factor under investigation, but in view of the small number of subjects in the study and the usefulness of the raw data analysis, further treatment of the data did not seem justified at this time This study was intended to be exploratory in nature and has served to indicate that additional work in the area might perhaps result in a number of useful concepts, whose practical application can contribute to the improvement of the nation s labor management relations

APPENDIX

Sample factual questions from the Labor Relations Information Inventory

(3) In the United States, organized labor comprises a) about 25 per cent of the labor force b) about 35 per cent of the labor force

(11) The monthly publication of the Bureau of Labor Statistics is the a) Labor Letter, b) Monthly Labor Review

Sample non factual' questions from the Labor Relations Information Inventory

(4) In 1947, the average weekly earnings in the bituminous coal industry amounted to a) $76 , b) $56 Correct answer $66

(5) At present, the following percentage of people in the United States are entirely dependent upon jobs and have very few savings a) about 55 per cent, b) about 85 per cent Correct answer about 70 per cent

(15) The recent increases of the price of steel are a) proportional to the wage gains made by the unions, b) proportion

ally greater than the wage gains made by the unions Meaningful answer not easily accessible

(24) In 1929, 49 per cent of the cor porate wealth in this nation (excluding insurance companies) was controlled by approximately a) 100 corporations b) 300 corporations Correct answer 200 corpo rations

(35) After one year of operation, the Taft Hartley Act has resulted in a trend a) toward successful management defenses of its rights and prerogatives, b) toward weakening the security of even the largest unions Correct answer controversial

REFERENCES

1 Breen, V E, *The United States Conciliation Service* Washington Catholic University Press 1943 139–150

2 Bullen, F H, The Mediation Process," in *Proceedings of New York University First Annual Conference on Labor* 1948, 105–143

3 Dalton M, The Industrial Rate Buster A Characterization,' *Applied Anthropology* 1948 Vol 7, 5–17

4 Gellhorn W and Brody, W, Supervisory Mediators Through Trial by Combat,' *Public Administration Review* 1948 259–266

5 Guilford Martin Personnel Inventory 1943, published by the Sheridan Supply Company, Box 837, Beverly Hills, California

6 Mediator Examination Announcement No 141, U S Civil Service Commission, Washington, D C, issued December 8, 1948

7 Warren, E L, and Bernstein, I, The Mediation Process, *Southern Economic Journal* 1949 15, Vol 4 441–458

8 Weschler, I R, An Investigation of Attitudes Toward Labor and Management by Means of the Error Choice Method" To be published in the *Journal of Social Psychology*

9 ———, 'A Follow up Study on the Measurement of Attitudes Toward Labor and Management To be published in the *Journal of Social Psychology*

10 ——— 'Who Should Be a Labor Mediator? *Personnel* November 1949, 222–228

11 Wherry, R J and Fryer, D H, Buddy Ratings Popularity Contest or Leadership Criteria? *Personnel Psychology* 1949 Vol 2, 147–160

12 Wonderlic Personnel Test, 1945, published by E F Wonderlic, Glencoe, Illinois

Chapter VI

MUSIC IN INDUSTRY

Although the presentation of music during hours of work is widespread, enthu siastic positive claims are not based upon sufficient evidence When business or ganizations pay for such services, it is safe to assume that they are not primarily interested in either entertaining their employees or raising their cultural level They are principally and reasonably interested in having the music affect pro duction favorably

This environmental change in working conditions readily lends itself to test by experiment Here is a situation in which the experimental factor can be varied in a deliberate and known manner, and changes in production can be measured Problems concerning the type of music played and the duration of intervals can be reasonably determined rather than guessed

This chapter is valuable because it illustrates the complexities of a problem that superficially seems simple Typical of most research, this problem reveals many difficulties in conducting controlled experimentation in industry

The reports of Kerr, Smith, and McGehee and Gardner, have been selected as

illustrative of work in this field Kerr's investigation concerned itself with deter mining the attitudes of employees regarding music scheduling The subjects had previously heard industrial music and so it was assumed that their opinions were related to their experiences The discrepancy between their attitudes and typical scheduling becomes obvious with examination of the data

Smith, in a monograph, reports the results of an investigation not only of attitudes but also of the relation between production and music scheduling His ex periment, despite the introduction of controls, displays no clear cut relation between music and accident rate, but does find that music is related to increased employee production and satisfaction Because of the length of the monograph only segments are reported

The McGehee and Gardner study does not find an increase in production oc curring with musical presentation Both the Smith study and this one are examples of sound industrial experimentation and so the discrepancy in results cannot be attributed to faulty procedure by one or the other The difference in findings may illustrate the generalization that music will increase production in simple industrial tasks, but not on more complex jobs McGehee and Gardner, however, offer the hypothesis that when a worker acquires a stable level of production, (for whatever reason), then music does not change the pattern This would then make the gen eralization subject to further experimentation before it would be acceptable

The work with industrial music is valuable insofar as it indicates the possibility of drawing faulty conclusions when the experimental approach is overlooked It shows that favorable attitudes no matter how overwhelming, cannot be regarded as evidence of actual effect This is a point worth remembering since it applies also to areas other than industrial music

Another noteworthy point concerning the relation between music and production is the demonstration that experiments can be conducted in industrial estab lishments without fear of dire consequences

Worker Attitudes Toward Scheduling of Industrial Music *

WILLARD A KERR

It is possible, though not necessarily true, that the average factory worker is the best authority on whether or not he should have music when he works, how much he should have, how he should have it, and when he should have it At least, his opin ions on these topics are important and should be investigated Already it has been demonstrated that factory workers want music (2), that music helps certain aspects of morale (1), and that music increases net worker output in monotonous opera tions (3)

Actual programming of music for fac tory audiences, with special reference to the time factor, now usually is done in one of the following ways by the plant broad casting director

1 *Fatigue dip periods* In some factory operations a temporary decline in output typically appears at about the middle of each half of the work spell Some plants schedule most or all of their music pro

* Reprinted from *Journal of Applied Psychology* Vol 30, No 6, December 1946

grams at these periods of believed fatigue and boredom

2 *Regular interval programs* Many plants set up a regular recorded music broadcast schedule which provides for 15, 20, or 30 minutes out of every hour of the work shift

3 *Employee request programs* A few plants do not follow a definite time schedule, but play records as they are requested by employees In one such plant the music, apparently well received, plays almost continuously

TABLE 25 1

Preference of 666 Factory Workers for Arrangement and Timing of Broadcast 'Music While you Work

	Per Cent of Employees Giving Each Response to Each of Three Major Timing Questions			
N =	224 Coil Winding	135 Phono Pressing	307 Tube Assembly	666 Total
1 *How much music do you want on your work floor?*				
A Never	00 5	00 0	00 6	00 5
B Lunch and rest periods only	01 0	00 8	00 0	00 5
C One hour out of eight	00 0	03 9	00 0	00 8
D Two hours out of eight	05 9	02 3	00 3	02 5
E Three hours out of eight	09 3	03 1	01 0	04 0
F Four hours out of eight	19 5	17 0	06 1	12 5
G Five hours out of eight	03 9	02 3	07 6	05 4
H Six hours out of eight	10 7	02 3	11 1	09 2
I Seven hours out of eight	01 5	03 1	09 6	05 7
J Eight hours out of eight	47 8	65 1	63 69	58 9
Total	100 0	100 0	100 0	100 0
2 *How do you want it?*				
A All in one session in the first half of shift	01 0	00 0	02 3	01 5
B All in one session in th second half of shift	00 5	00 0	00 3	00 3
C All in two sessions, one in the first half and one in second half of shift	10 2	09 3	07 2	08 5
D All in four sessions—one session of music in every two hours of work	21 8	18 5	15 3	18 0
E All in eight sessions—one session of music in every hour of work	35 4	23 2	30 0	30 6
F All in sixteen sessions—one session of music in every half hour of work	31 1	49 1	44 9	41 1
Total	100 0	100 0	100 0	100 0
3 *When do you want it?*				
A First hour	09 2	07 0	11 8	10 4
B Second hour	09 8	09 0	15 3	12 7
C Third hour	19 4	12 0	18 1	17 9
D Fourth hour	06 6	16 0	07 3	07 9
E Fifth hour	05 5	06 0	07 1	06 4
F Sixth hour	18 5	11 0	19 5	18 3
G Seventh hour	16 5	24 0	11 4	14 5
H Eighth hour	14 5	15 0	09 5	11 8
Total	100 0	100 0	100 0	100 0

While advocates of the various methods report favorable results, it seems that no attempt has been made to evaluate pref erences of workers for alternative methods in the time scheduling of music programs The average worker s opinion is a fact which must be regarded as important in evaluating the various methods, because the subjective fatigue boredom curve does not necessarily coincide with the familiar daily average hourly production curve, and factors other than fatigue and boredom may condition employees' time desires for music

Using the tear method of response de scribed elsewhere (4), a *Music Timing Ballot* was designed and administered to three groups of factory employees of the RCA Victor Division, Radio Corporation of America All were accustomed to work to music These 666 subjects represent a group of 79 females and 138 males en gaged in coil winding machine operations, plus a group of 99 females and 32 males engaged in pressing phonograph records in a Camden, New Jersey, factory, and 291 females and 7 males engaged in assembl ing radio tubes in a Harrison, New Jersey, plant Twenty failed to indicate sex Aver age age of the employees in the miscel laneous group is 30 6, in phonograph records 37 1, and in radio tubes 25 3 The miscellaneous and phonograph record manufacturing group heard a combination of the first two methods of programming mentioned above while the tubes group experienced the third method

Per cent of employees in each of the three groups giving a response to each question is indicated in Table 25 1 The responses to How much music do you want on your work floor?" tend toward bimodality although a majority of respondents, except in the miscellaneous group, indicate a desire for 8 hours of music out of an 8 hour work shift The average worker wants between 6 and 7 hours of music in 8 hours of work

A plurality of workers, if they were to receive 3 hours of music daily, want it di vided into 16 sessions, but the average worker wants approximately 10 sessions

In response to When do you want it?' a distinct tendency appears for the 2 middle hours of each half of the work shift to receive more votes than the first pre lunch, post lunch, or closing hour of the shift Music is least desired immedi ately before and immediately after lunch These subjective reports, probably based on feelings of fatigue and boredom, are particularly significant in view of the known tendency in many factory operations for output to decline temporarily toward the middle of each half of the work spell

Tetrachoric intercorrelations among the time variables sex, and age for all 666 subjects are shown in Table 25 2 Items one (how much) and two (number of sessions) come nearest of any 2 items to measuring the same thing, that is a general liking for industrial music and it is not surprising that the correlation between these 2 items is 66 Apparently morning or afternoon preference for music (Item 3) is not re lated with liking for music (Items 1 and 2) Older employees tend to care slightly less for industrial music and females seem to want more of it than do males It is true, however, that the mean age of the males reporting is significantly higher than that of the females Also, some older males tended to have jobs involving more super

TABLE 25 2

Tetrachoric Intercorrelations Among Five Items on the Music Timing Ballot for 666 Factory Workers

	2	*3*	*5*	*6*
1 How much	66	03	40	− 28
2 How (sessions)		00	26	− 27
3 When (afternoon)			− 06	02
5 Female sex				− 59
6 Age				

visory responsibilities These latter facts must be considered in interpreting the 2 following partial correlations Correlation of amount of music desired with sex when age if held constant by technique of partial correlation is 30, and a similar correlation of amount desired with age when sex is held constant is − 06 These results indicate that sex (female) more than age is a determinant of how much music a factory worker wants to hear while working, however, it again must be emphasized that sex in itself may be less of a real causal factor than the fact that work performed by the average male subject in this study is of a less monotonous nature than that performed by the average female employee

REFERENCES

1 Middleton, W C, Fay, P J, Kerr, W A, and Amft, F, The Effect of Music on Feelings of Restfulness—Tiredness and Pleasantness — Unpleasantness, *Journal of Psychology* 1944, Vol 17, 299–318

2 Kerr, W A, 'Three Studies in Plant Music, *Factory Management and Maintenance* 1943, *101* No 11, 280–286

3 Kerr, W A, 'Experiments on the Effects of Music on Factory Production *AAAP Monograph* 1945 Vol 5, 1–40

4 Kerr, W A, 'Where They Like to Work Work Place Preference of 228 Electrical Workers in Terms of Music,' *Journal of Applied Psychology,* 1943, Vol 27, 438–442

Music in Relation to Employee Attitudes, Piece-work, Production, and Industrial Accidents *

HENRY C SMITH

[EDITOR'S NOTE Only three excerpts from Smith's monograph have been selected for presentation in this volume 'The Music Program" describes the planned variation in musical presentation "The Use of Control Groups and Weighted Averages" indicates the value of controls and advantages of statistics as an aid in interpreting results "Summary and Conclusions' is presented to give the reader an idea of the scope of the problem investigated]

THE MUSIC PROGRAM

Music was begun on January 15 The amount, type, and distribution of music were varied in a predetermined fashion The recordings used in this study were drawn from the Radio Corporation of America's 'Economy Library" This library consists of 300 records which range from current popular selections to classics About 50 new records were added to this library during the 12 weeks of the study The collection, designed for industrial use, included mostly music of the 'hit parade" type although Negro, cowboy, patriotic, and more serious selections were also represented As all records were 10 inches in diameter, music time was based on 3 minutes for each record

The amount of music during the first eight weeks varied from 0 to 62 5 per cent of the working time—0, 1, 2, 3, 4, or 5 hours per shift The pattern of music was arranged so that each of the 8 weeks would have each of these 6 amounts of music once Thus, the pattern of music from Monday through Saturday of the first week was 2, 3, 0, 5, 1, and 4 hours per shift, the pattern for the second week, 5, 0, 3, 2, 4, and 1

The type and distribution of music for the varying amounts of music during the first 8 weeks were as follows

* Reprinted from *Applied Psychology Monographs,* No 14, 1947

No music One day each week for the 8 weeks

One hour of music One day each week for the 8 weeks Opened with 2 waltz or march selections The remaining 18 selections in different weeks were 0, 50, and 100 per cent vocals "Vocals' were all records in the Economy Library with any vocal portions The nonvocal selections were drawn at random from the rest of the collection The playing was evenly distributed throughout the 8 hours of the shift and began "on the hour' (e g, at 10 a m, not at 10 10 a m)

Two hours of music One day each week for the 8 weeks Opened with 5 waltz or march selections The remaining 35 selections in different weeks were 0, 50, and 100 per cent classics "Classics" were all records in the Economy Library which were relatively serious and complex, al though, as a group, they might more appropriately have been labeled "semi-classical" The nonclassical selections were drawn at random from the rest of the col lection The playing was evenly distributed throughout the 8 hours of the shift and began on the hour

Three hours of music One day each week for the 8 weeks Same type and dis tribution as for 2 hours

Four hours of music One day each week for the 8 weeks Opened with 10 waltz or march selections The remaining 70 records were selected at random from the library The 4 hours of music were evenly distrib uted in the first week, all in the first half of the shift the second week, and all in the second half of the shift the third week The distribution was similarly varied in the 5 weeks which followed

Five hours of music One day each week for the 8 weeks Same type and dis tribution as for 1 hour

The no music days constituted the "con trol group' The 1 and 5 hour days, it was hoped, would give not only an indi cation of the effects of these amounts of music, but also some suggestions as to the effectiveness of varying percentages of vocal selections The 2 and 3 hour days would give not only an indication of the effects of these amounts of music but also some suggestions as to the effectiveness of varying percentages of classical selections The 4 hour days would give not only an indication of the effects of this amount of music but also some suggestions as to the most effective distribution of music, and this could be further clarified in the music program for the last 4 weeks In order to test further the effect of the distribution of music, the schedule for the last 4 weeks included 5 days with 3 hours of music, evenly distributed throughout the shift, 5 days with 3 hours of music all in the first half of the shift, 5 days with 3 hours of music all in the second half of the shift, and 7 days with no music The selections during these 4 weeks were made from the library at the discretion of the operator according to no definite pattern Through out the 12 weeks of the experiment, the operator of the equipment on the day shift would lay out the records daily for each period during the shift The same records were played at the same period by the operators on the other 2 shifts

The Use of Control Groups and Weighted Averages

Under ideal conditions, all factors other than music which might influence produc tion would have been held constant throughout the 12 weeks of the study If this ideal had been achieved, any variations in production could then only be explained by variations in the music

In reality, numerous factors other than music were varying during the experiment and affecting production Figure 26 1(3A) shows a marked trend upward in produc tion over the 12 weeks of the study, each week through the experiment had approxi mately the same music so this increase in production was independent of music Fig ure 26 1(3B) shows a definite variation in production with the day of the week, each day of the week through the experiment had approximately the same music so this fluctuation in production was independent of music Figure 26 1(3C) shows a decided decrease in production efficiency with the total man hours worked, variations in the work force were unrelated to variations in music so this change in production was in dependent These changes in production

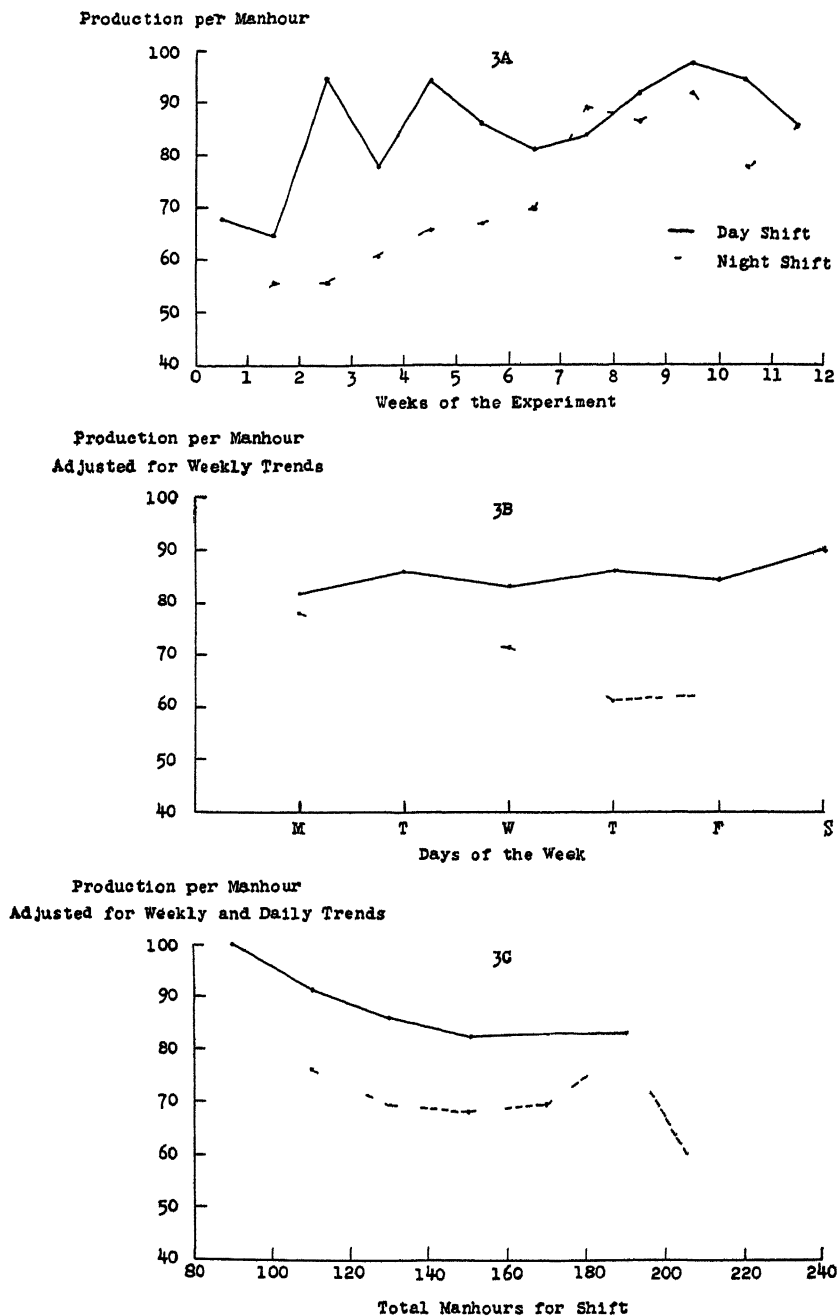



FIGURE 26 1 *Average production per man hour of rubber sealed terminals on the day and night shifts by weeks of the experiment by days of the week and by total man hours*

were due to such uncontrolled factors as the introduction of incentive pay, improved engineering methods, changes in supervision, better training, and individual motivational changes These were largely unrelated to music

The use of "no music days" in each week assisted in controlling most of these irrelevant factors By comparing production on no music days with production on music days, the influence of factors other than music tended to cancel out However, the changes in production due to nonmusic variables still tended to obscure the influence of music in two ways In the first place, they increased the standard deviation and thereby lowered the statistical significance of the differences between production under music conditions and under no music conditions Also, the nonmusic factors increased the correlation between production on the day shift and on the night shift, and thereby made the comparisons between shifts less significant because they were less independent of each other

To eliminate the variations shown in Figure 26 1 and the influence of the factors causing them, the average production per man hour of each shift was first computed for the 12 weeks of the study An index was then obtained for each week by dividing the average production per man hour for the 12 weeks by the average for the particular week Each day's average production within that particular week was then multiplied by this index to give the "Production per Man hour Adjusted for Weekly Trends" These figures were in turn adjusted in a similar way for variations due to the day of the week, and for variations due to the total hours worked The adjustments were made separately for the day and night shifts

Corrections for the hour of the day could not be made because production figures for the total line were not accurately kept by the hour Only the total production for the day was related to the accounting and payroll systems Individual records were kept only of the quantity of each employee's hourly production without regard to quality

The effect of the weighted averages may be visualized by reference to Figure 26 1 If music were not a factor of influence, each trend shown would become, with the adjustment, a straight line across the graph, parallel to the horizontal axis and at the average height of the line The statistical effects of the adjustments are shown in Table 26 1 The standard deviation of the daily average production per man hour decreased from 14 5 (day) and 15 2 (night) for the original data to 10 3 and 10 6 for the adjusted data As a result of these decreases the reliability of differences found in the study is increased The correlation between daily production per man hour on the day and night shifts decreased from 42 to 15 Comparisons between the day and night shifts were thus made of greater significance because of the relative independence of results on the two shifts

Thus, 3 types of controls were employed in the study Intra group controls were obtained by means of comparisons between

TABLE 26 1

Means and Standard Deviations of the Original and Adjusted Production per Man Hour for the 72 Days of the Experiment, with the Correlations between the Night and Day Shifts (N = 72)

Category	*Production per Man hour*			
	Original		Adjusted	
	Day	Night	Day	Night
Mean	82 0	63 9	83 0	64 5
Standard Deviation	14 5	15 2	10 3	10 6
Correlation	42 ± 10		15 ± 12	

no-music and music days Statistical con trols were introduced by means of weighted averages Intergroup controls were made possible by comparisons between the day and night shifts

The statistical unit used in the production study was the average hourly production for the day This unit figure was obtained by dividing the total production for the day by the total man hours worked At the maximum each unit is based on 168 man hours of work, 21 employees times 8 hours of work for each N' in the statistical formulae therefore is the number of days of production being considered

Summary and Conclusions

The influence of an industrial music program, which systematically varied the amount, type, and distribution of music played, was studied in a plant of approximately 1000 employees over a 12-week period An effort was made to determine effects of music on employee attitudes, piecework production, and industrial accidents

Employee attitudes A questionnaire concerned with attitudes toward music was sent to every employee before the music program began Somewhat over 70 per cent of these questionnaires were completed and returned An analysis of the replies showed that

1 Almost all of the employees (98 per cent) thought that music during working hours would be at least 'mildly pleasant," and 74 per cent thought that it would be "extremely pleasant"

2 The intensity of interest in music while working decreased somewhat with age The oldest group preferred semiclassical, nonvocal, and quiet music more than the younger groups

3 No sex differences in the intensity or type of musical interest were found

4 Personal interviews with a sample at the end of the 12 weeks showed no decrease in the desire for music while working

Piecework production Music in relation to production was studied on a highly repetitive assembly line operation which was on incentive pay Two separate shifts with an average of 21 employees on each shift were studied simultaneously for twelve weeks The results showed that

1 Production under varying conditions of music increased from 4 to 25 per cent The average increase on the day shift was 7, on the night shift, 17 per cent The increases were statistically significant and large enough to be of economic importance

2 Maximum production increases were found when music was played 12 per cent of the time on the day shift, 50 per cent of the time on the night shift

3 Production tended to decrease with a large increase in the number of semiclassical selections but did not vary with a large increase in the number of vocals Waltzes were more effective at the opening of the shift than marches

4 Production increases varied with the hour at which music was played and were greatest during the hours of low production

5 The more an employee wanted music, the more music tended to increase her production, the lower the employee s production, the more music tended to increase her production, the more the employee's job permitted conversations while working, the more music tended to increase her production

6 The greater effectiveness of large amounts of music on the night shift corresponded with a greater demand for music on the night shift the greater effectiveness of varied music corresponded with an expressed preference for varied rather than for special types of music, the greater effectiveness of certain distributions of music corresponded with an expressed preference for such distributions

Industrial accidents The 12 week music program was related to individual accident records kept for the entire plant The results of this analysis, none of which showed statistically significant differences, were as follows

1 No difference was found between the number of accidents on music days and on no music days when the three shifts were combined

2 The day shift had a relatively large

increase in the number of accidents with music, the other two shifts, slight decreases

3 Accidents tended to increase on the day shift with increases in the amount of music, but not on the other two shifts

4 Accidents tended to increase with a large increase in the number of semiclassical selections played, but not with a large increase in the number of vocal selections played

5 Accidents tended to increase more with waltzes in the opening period than with marches

6 Accidents tended to decrease with music in the earlier part of the shift but to increase with music in the latter part

Conclusions Music during working hours will generally improve production where repetitive work is common Properly administered in such situations, it not only will increase production but also will provide widespread employee satisfaction Music probably produces its major direct effect when the individual s capacity for attention is not absorbed by his work, in this circumstance, music appears to divert unused attention from brooding, talking, or off the job activities Although music, on the average, had no influence on the accident rate, the relation of music to accidents was not entirely clear in the present study

Music in a Complex Industrial Job *

WILLIAM MCGEHEE and JAMES E GARDNER

In Brief

The majority of published reports of the effects of music upon job performance of industrial workers are impressionistic and nonquantitative The few quantitative reports have indicated that music resulted both in an increase in the amount of production as well as an improvement in the workers attitudes towards their jobs This investigation is concerned both with the effects of music on the amount of production and with the way the workers believe the music affected their job performance The subjects in this study are women employees performing a complex industrial job in a relatively stable work situation The introduction of music in this situation resulted in no increase in production although the employees believed that music was beneficial to them in the performance of their jobs

The Problem

Industrial music has been used widely in American industry The exact extent of its use is not known but a recent report (4) estimates that there are as many as 6,000 industrial installations in the United States In spite of the wide use of industrial music there are few control studies of its effect on the workers and on the performance of their jobs Too often the effect of music on production, absenteeism, turnover, accident rates, and workers attitudes is measured in terms of the optimistic beliefs concerning its effectiveness held by those responsible for its installation and programming

The majority of investigators agree that the increase in production attributed to music comes not from rhythmic pacing but from the salutary effects music has on workers' attitudes For example, Smith writes, Music can increase production only through stimulating changes in the attitudes or behavior of the employees Improvements in production are by products of these changes" (3, p 54) It is implicit also in much writing on the subject of worker productivity that improved attitudes will result in increased worker productivity Following this line of reasoning, if music improves attitudes toward work, it should also increase production

* Reprinted from *Personnel Psychology,* Vol 2, No 4, Winter 1949

This study is designed to investigate certain aspects of the following problems

1 What is the effect of introducing music on the amount of production of industrial workers?

2 What are the opinions of the workers in regard to the effects of music on their work behavior?

3 What relations exist between the effect of music on production and the opinion of workers as to the effect of music on their work behavior?

The writers are aware that the results reported here may be peculiar to the work situation in which this study has been carried on

This study should be of interest also in that the conditions under which it was conducted differed from those surrounding two of the better controlled studies of industrial music Both Kerr s (2) and Smith's (3) studies involved workers performing relatively simple industrial tasks The work required of the subjects in this study is relatively complex as industrial jobs go Smith's investigation was conducted during a period of industrial expansion, employee training, and war enthusiasm The present investigation was made in a stable work situation All employees were experienced workers, production demands were stable, wages were based on an incentive system which had been in operation for several years without modification, there had been no change in supervision for three years In other words, the study was made in a relatively stable work situation on workers fully familiar with their jobs and reasonably well adjusted to the social environment of the work situation

Both Kerr and Smith reported increases in the amount of production associated with the use of music The former, it is true, based his conclusions on consistent rather than statistically reliable differences in favor of music The findings in this study concerning production are not in agreement with the results obtained by Kerr and Smith This is not to imply any inaccuracy in their work but rather to point up the fact that a different situation can lead to different results It may serve as an antidote to the practice of making generalizations from a sample of behavior to a universe

Subjects and Jobs

The subjects involved in this study are 142 women workers employed in the occupation known as 'setting' in rug manufacturing Eighty two of these operators worked on the first shift, 60 on the second shift The index of production used in this investigation is the average hourly output of these workers based on the units established by careful time study

The task of setting is a relatively complex industrial job It involves the preparation of material for rug looms The time required to reach the minimum skill where the worker is at a breakeven point between pay and production ranges from 6 to 15 months Two to four years of experience are required to become a skilled operator Without describing the job in detail, its complexity can be indicated by stating that skilled workers must have a high level of mental and manipulative skill, they must be able to attend to numerous job demands, they must possess high visual memory and color discrimination The job also requires considerable physical endurance since it involves constant standing and walking It also requires that the workers make an adjustment to a partner, since the work is performed by pairs of operators and pay is based on the output of the pair

Music and Production

The writers are indebted to both Smith and Kerr for the methods used in studying the effect of music on production in this study The effect of music on production was studied by the comparison of the amount of production on days on which music was played with the amount of production on no music days during an experimental 5 week period The following procedure was used in making this comparison The experimental period was begun one week after the installation of music This week was used for the purpose of ironing out difficulties in music equipment and in programming During each

of the 5 weeks of the experimental period, music was played four days and was not played one day

Four distinct music programs were used on each of the music days in a week The amount of music played each day during actual work hours was the same, 80 minutes The type of music played during the work hours, opening period, and lunch period was identical in each music program It followed recommendations for programming industrial music made by Benson (1) The programs differed, however, in the amount of music played in individual work music periods and in the use of music for opening, closing, and lunch periods

There were, therefore, four music programs (A, B, C, D) and one no music program (E) These programs were rotated during the 5 week period so that no one program was played on the same day twice This rotation was planned to minimize the effect on production arising from possible daily and weekly variations It allowed, also, for rigid statistical test of any differences found in terms of the effect of music, of no music, of weekly variations in production and of daily variations in production This test is described in the technical section of this report

Table 27 1 gives the arrangement of programs and the average hourly production by weeks, days shifts, and programs during the experimental period Table 27 2 gives the average hourly production for weeks, days and programs by shifts An analysis of these data shows no significant differences in production which can be attributed to any music program, to the lack of music, or to variations in weekly or daily production In other words, the only possible conclusion is that during the experimental period industrial music had neither a favorable nor unfavorable effect upon the production of these workers as a group It is possible that the production of the individual workers may have shown greater variation with music than without it These data, however, have not been analyzed to determine the nature of individual differences It is doubtful if this would be a fruitful procedure since the employees worked in pairs

Production considerations made it advisable not to extend the experimental period beyond 5 weeks It is possible that a longer period of adjustment might have been necessary before the maximum influence of music on production could be realized A comparison, however, was made between production during a 5 week period subsequent to the experimental period and production during a 5 week

TABLE 27 1

Average Hourly Production for Each Day in Terms of Weeks, Programs, and Shifts

Week	*Shift*	*Program*	*Monday Production*	*Program*	*Tuesday Production*	*Program*	*Wednesday Production*	*Program*	*Thursday Production*	*Program*	*Friday Production*
1	1	A	133	B	139	C	140	D	140	E	145
	2		123		114		120		124		111
2	1	B	136	C	141	D	143	E	146	A	139
	2		115		119		119		122		115
3	1	C	140	A	138	E	142	B	139	D	139
	2		114		116		121		120		107
4	1	D	129	E	132	A	137	C	136	B	140
	2		120		126		118		116		114
5	1	E	132	D	144	B	143	A	142	C	142
	2		114		122		123		124		118

TABLE 27 2

Average Hourly Production for Days, Weeks, Shifts, and Music Programs

	Weeks				
Shifts	1	2	3	4	5
1	139 4	141 0	139 6	134 8	142 0
2	118 4	118 0	115 6	118 8	120 2

	Days				
	Mondays	Tuesdays	Wednesdays	Thursdays	Fridays
1	136 0	138 8	140 6	140 7	140 4
2	117 2	119 4	120 5	120 1	113 0

	Programs						
	A	B	C	D	E	Music	No Music
1	138 0	139 0	139 0	139 0	141 0	139 0	141 0
2	119 0	117 0	117 0	118 0	119 0	118 0	119 0

period immediately preceding the installation of music While this procedure does not allow for careful statistical control of the effect of daily and weekly variation and other factors influencing production, it is interesting to note that the difference in average production between these two periods is not statistically significant The average hourly production (both shifts) for the 5 week period prior to music was 130 8 while for the 5 week period with music subsequent to the experimental period, average hourly production was 131 0 It seems unlikely, therefore, that an extension of the experimental period would have revealed any significant differences in production that could be attributed to the presence or absence of music

Employees' Opinions of the Music Program

Music, then, had no favorable or unfavorable effect upon the production of these workers as a group This failure to have any effect on production might be traceable to the employees like or dislike of the music program Accordingly, we developed a questionnaire to be administered to these workers to determine first, their general reaction to the music program and second, to determine how the employees felt that music affected their work The questions used in the questionnaire were based on data collected from intensive preliminary interviews with 14 members of this department These workers who were interviewed were selected at random

The questionnaire was administered to the entire group However, due to absenteeism and to a few returned questionnaires which were unusable, there were only 130 questionnaires which could be analyzed In other words, the results which are to be reported represent the opinion of slightly over 90 per cent of the workers in this department The results of this questionnaire indicate clearly that the workers were favorably disposed toward the music program In reply to a question "Do you want us to continue playing music in this department?,' 84 5 per cent answered in the affirmative Only 1 per cent answered in the negative, while 14 5 per cent indicated that it made no difference to them whether or not we continued the music

When asked about the specific aspects of the music programs such as the type of music used and its programming, the answers in the main were favorable The

major complaint that we received regarding the type of music played was that we were playing too much semi classical and Latin music, and not enough hymns

The music program, then, was received favorably by the majority of the workers Yet, as indicated above, it had no effect on the amount of production We, therefore, asked on the questionnaire the specific question, What effect does music have on your work?" The checklist which was submitted to the group was again based on our preliminary interviews Table 27 3 gives a summary of how music seemed to affect the job performance of these workers In general, music seemed to reduce monotony, to make time pass more rapidly, and to make the work easier It is interesting to note, further, that 59 per cent of the group said they got more work done with music as compared with a negative response of 7 per cent The opinion of the workers that music helps them to produce more is extremely interesting in view of the fact that there was no increase in measured production

As indicated, the rank and file employees in this department were favorably disposed toward the music program and felt that it helped their work We secured a similar reaction to the music from the supervisors This was secured through a questionnaire issued separately to supervisors at the same time we issued the questionnaire to the workers The items in this questionnaire were based again on information secured in informal interviews with these men

Returns were secured from the total supervisory force in the Setting Department, five foremen and assistant foremen In addition, returns were received from two supervisors in a small department adjacent to the Setting Department in which the same music was played Since it was

TABLE 27 3

Workers Respond to Question 'What Effect Does Music Have on Your Work? '—N = 130

Reported Effects	*Per Cent Responding*		
	Yes	No	Can t Tell
A Makes time pass	90	3	7
B Takes your mind off other things	74	14	12
C Gives you a lift when you re tired	86	4	10
D Makes you feel more like coming in	74	6	20
E If you come in feeling bad, music helps	82	5	13
F Music keeps work from getting on nerves	73	6	21
G Music gives you something to look forward to	75	5	19
H The hard patterns seem to come easier with music	49*	14	37
I You get more work done with music	59	7	34
J Music lets you know how much time has passed	65	6	28
K Music helps you know if you're behind or ahead in your work	49*	10	41
L You move in time with the music	54*	18	28
M Music breaks monotony	73	4	23
N You do less talking with music	80	5	15
O You seem to have more pep with music	77	6	17
P Interferes with your work	4	74	21
Q Makes you nervous	6	75	19

* Differences between percentage answering yes' and combined percentage answering 'no and can t tell are *not* statistically significant Remaining differences between yes and other responses are statistically significant

TABLE 27 4

Questionnaire on Effects of Music Administered to Supervisors and Number Replying (N = 7)

1 Do you want to continue playing music in your department?
Yes 7 No 0 Doesn t matter 0

2 Do you think the money spent on music is a worthwhile investment your department?
Yes 6 No 0 Possibly 1

3 Have the following improved or become worse since we started playing music in your department?

	Improved	*Become worse*	*No difference*	*Can t tell*
Employees gripes and complaints	5	0	0	2
Employees attitude towards Company	4	0	1	2
Employees attitude towards their work	5	0	1	1
Employees cooperativeness	4	0	3	0
Employees dependability	1	0	4	2
Quality of employees work	4	0	2	1
Employees relationship with supervisors	4	0	1	2
Employees relationship with each other	5	0	0	2
Your satisfaction in your own work	6	0	0	1
The general spirit in your department	7	0	0	0

desired to give these supervisors complete anonymity in making their replies to the questionnaire, no attempt was made to keep the two departments separate Table 27 4 gives a compilation of the responses from both the departments to the questionnaire

These responses indicate that the supervisors believed that the music improved employee attitudes, gave the employees' morale a lift, created better interpersonal relations, and increased job satisfaction among the supervisors themselves The supervisors believed further that music had made their duties easier to handle All of them wanted music continued, all except one of the supervisors believed that music was a worth while investment in their department In other words, these supervisors believed that music had definitely improved the attitude of their workers as well as their own attitude toward their job

Implications

As shown above, there is evidence that the employees in this study were favorably disposed toward music We have, further, evidence that they believed that music not only made the work more pleasant but, also that it increased their actual production On the other hand, within the limitation of the experimental design, we have evidence that there was no statistically significant change in the amount of production

As indicated earlier, it has been implicitly assumed by most writers on the problems of workers attitudes, that an improvement in the attitudes of workers and a reduction in the monotony of the task would tend to increase production Music, as a non financial incentive, is assumed to increase production by bringing about changes in the attitudes or behavior of the employees (3, p 54)

Smith further has suggested that "the more complex and varied the job, the less likely music is to increase production on it" (3, p 55) The reasoning here is that if the job is so complex that it requires the full attention of the employee, he will not attend to the music and the music will have no effect on his production It might also be implied that if the job is so complicated that it requires the full attention of an employee music might serve as a distracting element and thus reduce production

The task which the employees in this

study performed is a complex one It is possible that the failure of music, in this study, to increase production can be attributed to the fact that the entire attention of these workers was devoted to the job demands This does not seem, however, to be the case Both through observation and interviews with employees we have evidence that, in spite of the complexity of the task, workers were strongly aware of the music and that they sometimes hummed the tunes or sang the words of the music The answers to the questionnaire items themselves indicate that music received considerable attention from the workers It seems therefore, that the failure of music, in this study, to increase production cannot be attributed to the failure of the employees to attend to the music due to the complexity of their tasks Further, attention to music did not result in any significant loss in production

It seems to us, therefore, that an alternate hypothesis must be advanced to explain the failure of music within the limits of this study to increase or decrease production The workers in this study over a long period of time have reached relatively stable levels of production They have developed definite habit patterns of work and tempo of work Further, they have developed a fairly adequate adjustment to the social and task demand of this work situation It seems, therefore, in this very stable situation that the effect of music was not sufficiently strong in spite of its other salutary aspects to break up these well established habit patterns While some of the workers used music as a pacing device, i e , as a means of knowing whether they were behind or ahead in their work, the music did not change their production goals These workers, therefore seem after long experience on this complex job to have reached a stable level of production This level may have been arbitrarily established by the workers or may represent a physiological limit In any case, it is apparently so strongly established that it was not affected by music despite the favorable reactions of the workers toward the music Moreover, it seems that workers opinions regarding the effect of music on their own output cannot be taken as evidence of the actual effect

As indicated earlier, the production of individual workers may have been favorably or unfavorably affected by the music Due to the nature of the task, it is impossible to determine whether or not this

TABLE 27 5

Analysis of Variance Data and F Ratios Based on Production During Five Week Experimental Period

	First Shift			*Second Shift*		
Source	SS	DF	S^2	SS	DF	S^2
Programs	33 76	4	8 69	15 2	4	3 80
Weeks	97 76	4	24 44	66 0	4	16 50
Days	84 56	4	21 14	212 4	4	53 10
Error	165 48	12	13 79	206 4	12	17 20
Total	381 76	24		500 0	24	

F Ratios *

	First Shift	*Second Shift*
Programs × errors	0 630	0 221
Programs × weeks	0 356	0 230
Programs × days	0 411	0 072
Weeks × days	1 156	0 311
Weeks × error	1 772	0 959
Days × error	1 533	3 087

* None of the F Ratios is significant at the 5% level

is so. If it were so, favorable effects on one worker are masked or cancelled by unfavorable effects on other workers.

Technical Section

The programs of no music and music during the experimental 5-week period were so designed that the data could be analyzed by means of the Latin square analysis of variance technique. On the basis of the data in Table 27.1, we made a separate analysis for each shift. The analysis of variance data are shown in Table 27.5. In no instance is the F ratio statistically significant. This rigorous statistical test substantiates our earlier conclusions that production was not materially increased or decreased by the music, type of program, or absence of music during the experimental period.

We have tested the significance of differences between the percentages (Table 27.3) of those saying "yes," "no," and "can't tell" to the questionnaire items. In every instance except three, the differences are statistically significant between the percentage answering "yes" and the combined percentage for "no" and "can't tell." In these three instances the differences are statistically significant at the 1 per cent level between those who say "yes" and those who say "no."

Statistical treatment of our data, therefore, substantiates the conclusions drawn earlier in this study that music had no effect, favorable or otherwise, upon production of this group of workers. The workers, however, were favorably disposed toward the music and believed that it favorably affected their job performance. We have no experimental verification of our hypothesis, however, that music failed to affect production favorably in this situation due to the long established habit patterns of work in a stable work situation.

References

1. Benson, Barbara E., *Music and Sound Systems in Industry*. New York: McGraw-Hill Book Company, 1945.
2. Kerr, W. A., "Effects of Music on Factory Production," *Applied Psychology Monographs,* 1945, No. 5.
3. Smith, Henry C., "Music in Relation to Employee Attitudes, Piece-work Production, and Industrial Accidents." *Applied Psychology Monographs*, 1947, No. 14.
4. Spears, Ethel M., *Music in Industry*. National Industrial Conference Board, Studies in Personnel Policy, No. 78, 1947.

PART THREE

Engineering Psychology

Engineering psychology is a newer branch of industrial psychology Its prime concern is human problems related to machine design Most appropriately, it should not be considered as separate from industrial psychology, but as a part of it Its findings must ultimately be integrated with the findings in such other topics of industrial psychology as selection, training, work environment, motivation, and so forth

Whereas engineering is concerned with improving equipment from the point of view of mechanical design, engineering psychology is interested in adapting the equipment to man based upon his psychological capacities and limitations

It is one thing to design automobiles in accordance with principles of engineering and aesthetics but it may be that disregarding man's characteristics and limitations is contributory negligence if one meets death using the machine An automobile should allow maximum visibility In many cars, however, drivers cannot even see the fenders! Some insist this is unimportant since one can learn to drive by "feel" rather than by sight In addition, to discriminate among the gadgets on the dash-board is sometimes difficult, and operating some of them requires that the driver actually shift his entire body

With mechanical improvements in cars, problems do arise For example, cars with an automatic shift are no longer equipped with the foot clutch Then, rather than assuming that its present location is best, the position of the foot brake becomes a problem for the designer Other examples may be cited The hand brake no longer plays the role it formerly did, nevertheless it is still a useful part The handle by which the brake is operated varies from car to car and so does its location Since more people are right handed the question may be raised about the desirability of its location especially where it is necessary to use the left hand to operate it

The point raised is not that the hand and foot brakes are poorly located, but rather, are they most efficiently located in terms of man's characteristics and limitations? The suggestion is not to ask drivers but to test driver reaction in terms of behavior Accordingly, horn, lights, radio, head-room, seating arrangements, and so forth, can be designed functionally thereby leading to greater safety and efficiency of operation The reference to the automobile is suggestive of research in engineering psychology

Despite the newness of this field, research has been prolific As in the other sections of this book no attempt has been made to cover the entire field Four representative chapters have been included, namely, a program for human engineering, design of displays, design of controls, and visibility and legibility The

first chapter was included because the material it contains presents an outline of the subject matter and program In a new field such articles have real value The remaining three chapters represent topics in which considerable work has been done An area of importance not covered is the design and arrangement of work space It simply is impossible to include all topics and still keep this volume down to reasonable size

Chapter VII

A PROGRAM FOR ENGINEERING PSYCHOLOGY

Wars require the urgent recognition of problems and bring into focus critical needs created by these very problems Work performed under such forced pressure often results in contributions that otherwise would have been delayed (This is meant as a factual statement and is not intended as a value judgment for or against war)

Under such circumstances, psychologists during World War I devised group mental testing Industry later recognized the value of such mental testing During World War II psychologists attacked the problem of relating human abilities to equipment design This they did by integrating their findings concerning man's sensory, motor, perceptual, and intellectual abilities with their own experimental methods applied to the field of machine design

As long as man operates a machine, one must take cognizance of the machine in relation to man The automobile and the airplane are but two illustrations An auto can be built to ride smoothly at speeds faster than the speed at which man can react safely Planes can now travel so fast that they outrace their own sound Such performance raises human problems of preservation of limb, life, and the pursuit of happiness

The papers of Mead and Kappauf are especially relevant in that they emphasize the problems of equipment design in relation to man and also are important historically since they reveal the types of problems the psychologist must solve Mead indicates clearly the need for the wedding of engineers and psychologists He confines the field to mechanized tasks and prefers not to include human relations and its personnel and labor relations This book prefers not to accept such a prescribed limitation, not only because such a view has an undue engineering bias but also, and more important, because it leads to an incomplete treatment of other equally practical problems It recognizes the importance of engineering psychology but only as one aspect of industrial psychology

Kappauf presents the view that increased efficiency is achieved by having the same "research team" continue beyond the problem of design and operation of equipment and into the problems of training the operating personnel

Dunlap is concerned with the problem of adapting machines to man and is aware that this task has wide application in such diverse fields as transportation, manufacturing, and even farming His article, broader in perspective than the previous two, draws less upon military illustrations and demonstrates how problems originally arising from military needs soon have transfer value to industrial needs

A Program of Human Engineering *

LEONARD C MEAD

The opinions or assertions contained herein are the private ones of the writer and are not to be construed as official or reflecting the views of the Navy Department or the naval service at large This paper, presented before the New York Academy of Science May 10 1948, is published with their permission

SUMMARY

A resume of the manner in which the field of 'human engineering arose is presented together with a specification of what is encompassed by the term This is followed by an outline of major topics of the field and representative experiments for each topic These researches indicate that human engineering is not a science or field of endeavor unto itself Rather it is a technical service which our society now demands Recent interest in the phrase merely indicates that individuals from many professions, particularly the engineering and biological sciences, have discovered that they can and should be of mutual assistance in solving the problems of fitting man to the complexities of his contemporary work situation

HISTORICAL PERSPECTIVE

The program to be discussed is not an ideal program which might be proposed to the world at large Rather, it is the one being sponsored at present by the Special Devices Center of the Office of Naval Research Like most programs it came into being gradually in order to meet various needs, emergencies, requests and other administrative necessities We are attempting to develop a logically complete and rounded set of projects so that in time it will be possible to describe the program and to show how each paper of a symposium such as this is related to the whole program As will be pointed out later on, we have a scaffolding fairly well set but a great deal of construction on the building still remains to be done

Here is the plan of presentation First, a few words of historical perspective to show the antecedents of the human engineering field Then, a definition of this field to distinguish it from related areas of endeavor Next, an outline of the program with which I am most familiar, and finally an illustration of some results already obtained from human engineering research

Human engineering problems seem to arise whenever man is confronted with technological advancements Perhaps the primitive cave dweller would have profited if his tools and weapons had been shaped and weighted so as to fit his psycho physiological capacities The need for such modifications, however, does not seem to have been very pressing, it resembles giving Aristotle a telephone Not until the machine age was well advanced in the 19th century did anyone do any systematic worrying about the fact that man was actually the weak sister in mechanized production The time study work of Taylor in the 80 s, followed by the motion study contribution of the Gilbreths, represent the first organized attempts to make a man a more efficient partner in the modern industrial scene

During this same period the notion was being developed in the new field of psychology that individuals are constituted differently and that some people are naturally better suited than others for particular types of jobs The mental testing movement led to the development of intelligence and aptitude tests World War I gave both impetus and status to personnel selection Thus, during the early

* Reprinted from *Personnel Psychology,* Vol 1, No 3, Autumn 1948

decades of the 20th century we find both engineers and psychologists attempting to adapt human beings to the demands of a technological society

Meanwhile another group of scientists, the experimental psychologists, were developing their field by the discovery of new facts and techniques concerning man's sensory and perceptual processes It was a long time before this group became enmeshed in the practical problems of matching man with the technical appurtenances of his civilization In fact, the leading experimentalists took pride in divorcing themselves from the applied aspects of their research and a number of psychological publications made the point that this psychology must remain "pure The stress of a World War II, however, brought this group of specialists into the human engineering field Early in the war it was observed that the potentialities of modern weapons and equipment, despite superior engineering achievements, were not yielding the performance that their advanced design seemed to merit It was recognized, too late in many instances, that the human operator remained as an essential link in military tasks and procedures, and that special effort is required to design the equipment so that it fits the natural characteristics of the average man It was the experimental psychologist who seemed to know the most about these normal capabilities of man What World War I did for the mental testing movement, World War II seems to have done for experimental psychology In tracing the history of this movement, it is interesting to note that engineers first invaded the domain of psychology by recommending certain behavioral procedures on the basis of their time and motion studies Psychologists have now returned the favor by specifying the manner in which equipment should be designed

Definition of Human Engineering

Our definition of the field goes like this human engineering is that endeavor which seeks to match human beings with modern machines so that their combined output will be comfortable, safe and more efficient Obviously, this kind of effort is not specific to any one professional group but rather requires the special aptitudes of many professions and individuals In addition to the engineers who devise a particular type of equipment there may be the need in some cases for motion and time study men, physicians, psychologists, physiologists or other specialists from the field of the biological sciences The problems of human engineering present great diversity Many industrial situations involve particular environmental problems of heat, noise, lighting, humidity, noxious gases and so forth Modern aircraft, both commercial and military, repeat and add to this list of the physical factors which may affect the human being adversely Military equipment has confronted us with the necessity for the design of instrument indicators which can be quickly interpreted without error and controls which are conveniently arranged and have physical characteristics which match the muscular capabilities of the operator The design and use of prosthetic devices is another area which has brought the engineer, medical man and psychologist together

The examples of human engineering endeavor just cited seem to constitute a unique and coherent set of problems Some individuals prefer to call this general area "engineering psychology," a phrase which is acceptable to the writer However, the line must be drawn somewhere It is proposed that we do not attempt to include in this field the problems of human relations, personnel management or labor relations The term human engineering" has been applied to all three of these fields during the past few months There is no objection, naturally, to these types of activity, but it is believed that the term human engineering will lose much of its value if it is applied indiscriminately to all attempts to fit the individual into his social and economic, as well as his machine, environment Let us confine ourselves, therefore, to that area of scientific endeavor which seeks an optimal rapprochement between the human individual and the mechanized tasks which he is required to perform in our society Some

are undoubtedly of the opinion that this laboring over the matter of definition is merely a bit of definitional hair splitting As one from a very small number of people who are actually called human engineers, and one who sincerely believes that the events of the past few years have produced a unique opportunity and need for the engineering and biological professions to work together, I am convinced that the recognition and delimitation of this biomechanical field is essential to its success

A Program of Human Engineering

The writer's experience is primarily with military problems and, naturally, the general trends and outlines which are to be mentioned stem from this experience Obviously, there are parallel phenomena and problems in industrial civilian life The application of military research results to such industrial problems will be apparent

The proposed definition of human engineering gave as a goal the man machine combination which was efficient, comfortable and safe So far as efficiency is concerned, the objective is to have a given operation accurate, rapid and without error One way of eliminating human error is to automatize the operation and thereby eliminate the operating personnel This commendable objective frequently backfires, however, by leading to a situation where some men are eliminated but others are left with tasks which are more complicated than ever Meanwhile our technology continues to produce machines, gadgets, equipment and vehicles which require of the user novel and difficult skills Since man does not undergo evolutionary changes as rapidly as the field of engineering advances, it behooves us to persuade the engineer to design his contributions so as to take account of what the normal individual is naturally fitted to do

Thus the first requirement of a human engineering program is the acquisition of information about man's natural capabilities and limitations Although a great deal of this knowledge has already been provided by the biological sciences, most of it was not gathered with our present objectives in view There is the need, therefore, for a careful review and analysis of psychophysiological data already in existence in the professional journals, service and OSRD reports so that those who are concerned with equipment design problems may use the knowledge which already has been gathered Because of the interdisciplinary skills which human engineering engenders, this kind of literature survey is regarded as of about equal importance with new research efforts

The types of experimental studies which also are required in an ideal program may be grouped as follows

Studies of the optimal environment Investigations are needed to specify the effects of physical factors having deleterious or favorable influence on operator behavior Great progress has been made in recent years by the engineering profession on the manner in which temperature, humidity and ventilation interact to affect employee productivity Other variables which are encountered in industrial situations are noxious gases, noise (sonic and ultrasonic), vibration and illumination Commercial aviation has given us the additional complications of oxygen deficiency, air pressure, acceleration and motion (8) Military circumstances make it even more difficult to obviate the undesirable effects of these physical factors Continued cooperation between members of the engineering, medical and psychological sciences will be required to protect the human individual while working in environments which offer hazards to his safety and efficiency

Studies of equipment display This area has been under attack for an appreciable period of time and many general principles are now available for application As these data become more widely known we can anticipate more functional designs of dials, scales, meters graphs and other types of instrument and tabular indicators Included in this problem of the display of machine information is the layout and arrangement of the working place Frequently a simple rearrangement or redesign of the machine indicators and layout greatly simplifies the sensory requirements of the job

Studies of equipment controls This is an area which has been neglected by psychological and medical investigators Present information has come from the engineering side of the ledger and is expressed in the general principles of motion and time study The rules of motion economy have been appraised by the practical yardstick of greater and faster production But is this the best criterion for making recommendations on man s motor performance? The old controversy as to whether there is one best way" to perform an operation still needs to be resolved Fortunately the cooperative efforts, begun during World War II, are now in progress on the physical design of machine controls In the near future we may expect definitive recommendations for the optimal size shape gearing, direction of motion, speed of motion, inertia, and friction of different types of controls in a variety of machine situations The problem of control design is associated closely with the specification of the ideal controls to be combined with the equipment display An important consideration in this regard is the extent to which a control knob wheel or lever can serve as an instrument display and provide additional information to the operator More data are needed on the extent to which the muscle sense can be utilized in keeping the operator informed of the progress of the task and the conditions of the machine

Studies of man machine systems This area is probably not as significant in civilian industrial situations as it is in military operations In the latter case there are many instances where large groups of men must coordinate their individual efforts so as to meet a single common objective Many instances of crew coordination are demanded in the control of vehicles such as ships and airplanes Communication networks also illustrate the fact that the human factor may determine whether a complex physical system will or will not function properly

In summary, the first requirement of a program of human engineering is the acquisition of knowledge which will specify what the normal working man can do naturally and effectively This objective can be attained by the collection and dissemination of published facts and the conduct of further research on the working environment, equipment display machine controls and man machine systems Much valuable information can be found in the stockpile of knowledge possessed by the engineering profession and the biological scientists Without cooperative effort on the part of these groups the solution to many present day problems will remain one sided or unsatisfactory

While speaking of the requirement for interdisciplinary effort, a further detail of explanation is appropriate Dr McFarland recently informed the writer of his experience during a meeting of engineers, physicians and psychologists who were discussing problems of aeronautics The medical men and the psychologists spent considerable time in announcing what was wrong about cockpit design from the point of view of physical safety and ease of operation To this some engineers took offense and pointed out, justly, that they made their airplanes according to design specifications furnished by others The point is that the engineers are not basically responsible for machine designs which neglect biomechanical considerations As suggested above, the ideal human engineering program will not only gather the basic data but will also disseminate them to those who can use them The study of the human being is within the province of biological investigators and they must accept the responsibility of translating this knowledge into a useful form and then cooperate with the engineering profession in its application A reversal of this experience has occurred in the development of prosthetic devices In this field it was found that the medical men were in dire need of engineering information This is merely another instance wherein professional teamwork is required to satisfy the needs Group endeavor implies that each member will contribute his special skill for common benefit At the same time each partner should attempt to understand the contributions of the other members of the diverse professional group

In concluding the presentation of a human engineering program, there remains

one additional item Thus far the discussion has centered about the modification of equipment, machines, vehicles, and prosthetic devices to take account of human characteristics An alternative procedure is to modify the individual through training procedures and devices The best human engineering endeavors will undoubtedly still leave *homo sapiens* in behavior situations which are complex and difficult Therefore, to round out an adequate program, the writer believes that personnel training is a part, or at least a necessary adjunct, of the total project

Some Recent Findings in Human Engineering

In this section of the paper some results of specific research investigations will be cited to illustrate the kind of information that will be forthcoming as the program outlined above becomes effective

The first area of problems mentioned was the fitting of the individual into atypical environments A most striking example of this type of problem has been brought about by the development of high speed, piloted aircraft It is difficult to imagine the visual, acoustic and vibrational environment for the pilot who flies straight and level at more than 600 miles per hour Insofar as he uses visual contact reference, his whole visual field moves at much greater rates than in traditional flights so that he approaches the limits of ordinary reaction time in observing and responding to stimuli What was a mild bump at 160 miles per hour now becomes a violent jar so that the pilot must wear crash gear to keep from being knocked unconscious If the pilot wants to do a turn he is now in danger of blacking out due to angular acceleration

In order to study the effects of centrifugal acceleration and to develop protective measures, there are several human centrifuges now in operation Data from studies employing a centrifuge show that when the pilot is exposed to a force of no more than 2 *g* in the direction of head to feet there is a marked feeling of pressure as he is forced into his seat and that the extremities become difficult to lift (1) Response time is increased accordingly At 3–4 *g* the heaviness of the extremities is exaggerated and great effort is required to move the hands and feet, erect posture is maintained with difficulty Between 5 and 8 *g* unconsciousness or coma develops, this state is preceded by blacking out of the field of vision, probably due to the loss of blood from the head and face The pilot s value in controlling his machine is practically nil at this point Yet the accelerative forces about which we have been speaking are well within the stress limits of the aircraft structures Whether it will be possible to bring the man s tolerance up to the tolerance limits of his aircraft is difficult to say It may be that here we have a true instance of engineering possibility not being realized because of human weakness (8, 10)

Illustrative research in the field of instrument displays will now be cited A large percentage of visual indicators are of the clock face type A question which faces the designer of such instruments is the number of graduation marks to put around the scale One might guess that it would be extremely easy to ascertain the answer to this question As a matter of fact, it has taken a number of years and several comprehensive experiments to begin to see a general solution to this problem

One of the first experiments was done during World War II by Loucks (7) who used aircraft tachometer dials with various numbers of markings On the basis of short exposures, he showed that the percentage of errors was greatest for the dial with the largest number of graduations He concluded that the cleanest dial from the standpoint of design gave the most accurate readings Grether and Williams (5) measured speed and accuracy of reading dials ranging from 1 to 4 inches in diameter and from 5 to 40 degrees in angular separation of graduations They found an increase in accuracy of dial reading as the dial diameter increased, except for the case with 40 degrees angular separation of graduation marks In the latter case there was a decrease in accuracy when dial diameter exceeded 2 inches They then plotted all their data on a single curve re

lating accuracy to length (*not* degrees) of graduation interval Regardless of diameter of the dial and the angular separation between graduation marks, the error was found to decrease as the linear distance between intervals increased up to ¼ inch with little improvement thereafter The most recent relevant experiment is that of Kappauf (6) who employed a different method of measurement and found, contrary to Loucks (7), that accuracy increased with an increase in the number of graduation marks A 5 unit dial was better in terms of error and speed of reading than a 10 unit dial, a 1 unit dial was only slightly better than the 5 unit dial

Grether (4) explains these diverse findings by noting that quantitative dial reading errors may be of two kinds, interpretation errors and interpolation errors Interpretation errors would be *increased* by an increase in the number of graduations because of the greater ambiguity as to which of the markings the pointer is over Interpolation errors, on the other hand, are *decreased* by an increment in number of dial markings because the reader has less difficulty in deciding how far along the pointer is between the scale divisions

The third major area of human engineering is the specification of the characteristics of machine controls One of the most significant problems here is the basic motor capacity of the average individual The Therblig notation system of the motion and time study engineer has long been a standard method for classifying the different elements of work performance In 1947, however, Brown and Jenkins (2) proposed a new classification which may serve as an impetus for further research In brief, they separate motor reactions into three distinct classes

1) Static reactions, which include all instances where a body member is required to be held in a fixed position in space,

2) Positioning reactions, wherein the members of the body are moved from a position of rest to a specified position in space, the terminal accuracy being of primary significance, and

3) Movement reactions, which are movements of the bodily members at given rates, in given directions, along specific paths

Subsequent to this analysis of motor reactions, Brown (3, 11) completed several investigations on discrete and positioning responses Mention of some of these results is warranted both because of their novelty and their significance One of these studies (11) was to ascertain the accuracy with which individuals could perform positioning reactions in the absence of visual corrective cues The subjects were required to move the right arm and hand from a point of rest to a terminal position located either 0 6, 2 5, 10 or 40 cm distant After an exposure of 2 5 seconds to one of these four extents the reactions were made in total darkness Movements were made in both horizontal and vertical planes in various directions It was found that there was a tendency to overshoot the intended mark at shorter distances and to fall short at longer distances One exception, attributed to the effects of gravity, was noted for all distances in the vertical plane when the direction of movement was downward The per cent error decreased and the variability increased with each increment in distance A plot relating speed of movement to distance was found to be of the exponential form $y = ax^b$

Because of the possible significance of speed of bodily movements to equipment design and job performance, Brown (3) has just completed a follow up study to determine the effects of speed up instructions on positioning reactions The motions were all in the horizontal plane Despite the emphasis on speed in the instructions to the subjects, there was no increase in the average reaction time Although primary movement time was decreased, the total movement time was not There was an increase in the time spent in making the fine, secondary adjustments following the initial gross approaches to the terminal point Brown concludes that attempts to speed one's movements may produce apparent increases in speed which, in terms of over all efficiency, yield little genuine improvement

The last area in the program of biomechanics was designated as the study of man machine systems (9) Rather than

present an illustrative sample of results obtained in this field, it seems preferable to outline briefly a method of attack on such problems which has been proposed by the Systems Research Laboratory of the Johns Hopkins University A paper by Dr Chapanis represents one phase of the Hopkins attack on the over all efficiency of man machine combinations Method ology is one of the main determinants of progress and this project's contributions to methods are of far reaching significance

The objective of psychophysical systems analysis is to specify (1) the most efficient number of human operators, (2) the num ber and characteristics of the equipment components, and (3) the best arrangement and layout of the men and their gear

To say unequivocally that such appraisals can now be made would be to exaggerate the facts What has actually been accom plished is the formulation of a systematic approach, that is, a theoretical scheme for attacking problems of this nature

The tentative nature of this theoretical structure prevents presentation of more than a skeleton outline of the procedure The first step is to analyze and itemize all the connections between all of the com ponents of the system, thus all connections between men and machines, men and men, and machines and machines are listed These connections, which are termed links, will be found in a majority of cases to be visual, 'auditory" or con trol in nature After determining what these 'links" are it is then necessary to estimate the importance of each one There are two criteria for determining impor tance or 'link value One is the fre quency' with which each link is used and the other is the importance of the link when it is used In the case of a system which already exists, the link value as measured by frequency can be deter mined merely by tabulating the number of times that a particular link is employed For determining the link "importance' value measures of a more qualitative sort involving psychological rating scale tech niques must be used These techniques may be applied to existing systems or to proposed systems not yet constructed When link use value' and 'link impor tance value" of an existing system have been estimated, the two measures are then combined into a single score which is used in the final rearrangement of the whole system This last step consists in arranging the over all link values in order of size and importance By using a graphical plot and juggling the link values around in proper manner, a proposed solution to the particular systems problem is obtained

This approach to systems design is ad mittedly qualitative in some respects and not rigorously scientific Whether or not we have here the essence of a basic and valid theoretical construct is not yet clear On the other hand this approach has been used in a number of military circumstances and has been found to increase the over all efficiency of complicated men machine linkages Further research and application by motion and time engineers and psychol ogists is needed to demonstrate both its generality and its limitations

References

1 Armstrong H G, *Principles and Prac tice of Aviation Medicine* New York Houghton Mifflin 1946

2 Brown J S, and Jenkins, W O, 'An Analysis of Human Motor Abilities Re lated to the Design of Equipment and a Suggested Program of Research In Fitts, P M (Ed) *Psychological Re search on Equipment Design* Report No 19, AAF Aviation Psychology Program Research Reports, U S Govt Printing Office, 1947

3 Brown, J S, and Slater Hammel, A T The Effect of Speed up Instructions Upon the Performance of Discrete Movements in the Horizontal Plane" ONR Report No 57 2 3, 1948

4 Grether, W F, The Design of Instru ment Dials for Ease of Reading Paper presented before SAE National Aero nautic and Air Transport Meeting, April, 1948

5 Grether, W F and Williams, A C, Jr, Speed and Accuracy of Dial Read ing as a Function of Dial Diameter and Angular Separation of Scale Di visions" In Fitts, P M (Ed) *Psycho logical Research on Equipment De sign* Report No 19, AAF Aviation Psychology Program Reports, U S Govt Printing Office, 1947

6 Kappauf, W E, Design of Instrument Dials for Maximum Legibility I Development of Methodology and Some Preliminary Results, USAF AMC, Aero Medical Laboratory Memorandum Report No TSEAA 694 1L, 1947
7 Loucks, R B, Legibility of Aircraft Instrument Dials A Further Investigation of the Relative Legibility of Tachometer Dials, AAF School of Aviation Medicine Randolph Field, Texas Project No 265, October 1944
8 McFarland, R A *Human Factors in Air Transport Design* New York McGraw Hill, 1946
9 Mead L C, Human Factors in Engineering Design, *Journal of the Society of Automotive Engineers* 1947, Vol 55, No 12 40–46
10 Mead, L C, Application of Human Engineering to Flight Problems *Journal of Aviation Medicine* 1948, Vol 19, 45–51
11 Slater Hammel A T and Brown, J S, Discrete Movements in the Horizontal Plane as a Function of their Length and Direction, ONR Report No 57 2 2, 1947

History of Psychological Studies of the Design and Operation of Equipment[*]

WILLIAM E KAPPAUF

When the Army and Navy recruited psychologists early in the war, assistance was sought primarily in the areas of selection and training of personnel Accordingly a great number of psychologists became engaged in programs which, as described in part in previous issues of this journal, involved the validation of selection and classification tests, the preparation and validation of various types of training aids, and the coordination of new tests or training procedures with those already in use

At the same time a few research programs sought the services of psychologists to insure the more satisfactory design of some items of military equipment, to insure design which would take account of particular psychological and physiological characteristics of human operators Typical of this work was that which was initiated in the design of dark adaptation goggles, sun scanning devices, and communications equipment As the war progressed this phase of research in applied psychology assumed greater and greater importance and involved more and more types of equipment The field developed as much through the initiative of individual psychologists as it did through specific service requests In many cases psychologists who had been requested to prepare new training materials or training bulletins found previous operating instructions incomplete or unstandardized Specific study and research was required to elaborate the old procedures or to demonstrate the relative merits of alternative methods of operation With service approval, psychologists conducted such research and then incorporated the indicated procedural changes in the training bulletins or training programs which they were developing But having extended the field of their work from how to teach' to what to teach,' many psychologists found it inevitable that their thinking turned to equipment design as it related to efficient operation Increasingly, contributions were made to the solution of design problems The effectiveness of these early studies of design led the services to request their continuation and extension

Of course, none of these areas—selection, training, equipment design and operation—was distinctly new to the psychologist These had been central problems in the field of industrial psychology for many

* Reprinted from *The American Psychologist* Vol 2, No 3, March 1947

years What was new, however, was the more general acceptance of the principle that operating procedures and equipment design should be established on the basis of sound psychological data

In number, the military problems of design and operation were many This was most certainly the result of the recent rapid strides which had been made in the technological design of equipment Engineering developments had outstripped the rate at which the engineer could adapt his equipment to efficient human use To the psychologist the field was a rich and challenging one As evidence of his interest one may note the many names he applied to describe his work *human engineering biomechanics, psychological problems in equipment design the human factor in equipment design, applied psychophysics systems research* Further evidence of the importance which participating psychologists attached to this work is found in the fact that special sessions were given over to problems of equipment design and operation at two major meetings of psychologists immediately after the close of the war [1]

Studies of operation and design problems typically followed a pattern familiar to those acquainted with similar industrial work Some problems were handled on a thorough going research basis until the best procedure or the best design and arrangement had been experimentally determined Others were handled as adequately as possible on the basis of facts or principles already available in the psychological and physiological literature Although there was no fundamental difference in approach or method between military and industrial research, certain distinguishing features of military job situations changed the emphasis or plan of the work to an appreciable extent At least two of these are worthy of mention

In the first place, a high proportion of military jobs are such that continuous precision or continuously acceptable performance is required These jobs resemble laboratory pursuit meter tasks but have the consequence that failure to maintain continuously precise operation may mean loss of life or failure to accomplish a mission In this respect, many military jobs have few or no counterparts in industry where quality or precision often reflects only a worker s finishing touches or his final skill in bringing a piece of work within a required tolerance In radar scope interpretation, tracking a target, and other military jobs, it is the continuity of performance that is important This makes it necessary for the psychologist to pay particular attention to studies of changes in job performance over very short as well as over long time periods

In the second place, the demand for speed of operation is extremely pressing in military work For this reason speed becomes a primary criterion in evaluating operating procedures and design arrangements It is necessary that all operational shortcuts which do not interfere with the maintenance of continuously accurate performance be perfected and made standard practice

Another feature of the war research program which psychologists found different from the usual work in industry was that they were invited with increasing regularity to participate in discussions and tests of equipment which was not yet in production but which was still in preliminary design or pre production form Increasingly, even though slowly, their jobs changed from that of doctoring or rearranging old equipment so that it might be operated with greater success, to that of constructive criticism and study of new devices and weapons This trend produced a real increase in the efficiency of the production and testing programs Design from both the technical and operational points of view was considered in a coordinated manner To be regretted only is the fact that this approach to equipment development was not achieved at an earlier date and with reference to more types of equipment

Projects concerned with psychological research on equipment were organized under the Applied Psychology Panel and under a number of other divisions of

[1] Joint Army Navy OSRD Conference on Psychological Problems in Military Training—August 15–16, 1946 Meeting of the Military Psychology Section of the APA—November 27–28, 1946

NDRC The specific directives of these projects varied but some of them were sufficiently broad to include a full range of problems—research on the design of equipment, research on procedures for operating equipment, the development of selection and training procedures, the preparation of training materials in the form of pamphlets and training aids, and, when expedient, the initiation of more fundamental research on psychological factors basic to equipment design and use General directives of this sort made it possible for projects to organize broad programs of research directed at all aspects of the operating and training job

In an enviable position, so far as research opportunities were concerned, were those few projects to which newly engineered equipment was made available for study This equipment was often provided for testing just as soon as a development laboratory or a manufacturer had turned out two or three pilot models Typical of these projects was one with which the author was associated[2] This project was organized to investigate psychological problems in the design and operation of new antiaircraft lead computing gun sights and gun directors The program was sponsored jointly by the Office of the Commander in Chief, the Bureau of Ordnance, and the Ordnance and Gunnery Schools at the Navy Yard, Washington, D C Through this combined sponsorship, the project enjoyed the strongest support and assistance A continuous personal and first hand interest in all aspects of the equipment problem was shared by the several members of the sponsoring groups

The project was set up in association with the Ordnance and Gunnery Schools One section of these schools had the responsibility of training maintenance personnel for fire control equipment At the time that the project was organized, this school was designated to receive at least one experimental model and at least one pre production model of each new fire control device These instruments were regularly available to the research project for study and test Although the amount of research time [illegible] the equipment and the number of subj[illegible] that could be obtained were usually limited, the work of the project had the important advantage of timeliness Team operating procedures were tested and developed on the experimental model of each device Teams were drilled and timed in the use of these procedures Design inadequacies which hampered swift and efficient operation were made the subjects of reports to the cognizant Navy groups In most cases these design comments were submitted at such a time that they could be given full consideration by production engineers before the device was put in final production form When a more detailed investigation of some design problem seemed necessary, the research was planned jointly by the project and liaison groups and then undertaken by the project

Two steps were taken to insure the indoctrination of Naval personnel in effective methods of operating the new equipment Under the direction of the Office of the Commander in Chief, the project supervised the instruction and drill of special training teams These teams learned the operating procedures which had been developed through project study They were then assigned to special duty at training stations or with fleet units where they were responsible for training new crews in the approved techniques To supplement and assist the work of the training teams, the project, in cooperation with its liaison groups, prepared pamphlets describing the best operating procedures and outlining practical methods of training men as operators

The specific accomplishments of the research project during its year and a half of work included an evaluation of the design and operating characteristics of eight different gun director systems, the preparation of detailed operating procedures for six of these systems, the study of several synthetic trainers intended for use with these directors, and the investigation of a number of psychological problems relative to the use of these systems The

[2] Project N-111, Applied Psychology Panel Personnel included Henry Birmingham, Clarence Graham, Thomas Hermans, Milton Horowitz, Alston Householder, William Kappauf, William Lambert, Henry Meyer, Franklin Taylor

latter included a study of methods of rating operator performance on the equipment, an analysis of operator learning curves under different tracking conditions, an experiment on the design of tracking reticles, and a determination of the accuracy of unaided visual range estimation on aerial targets

Now that peacetime research programs are being set up by the services, one may well inquire into the lessons learned from the work and organization of projects like the one just referred to The two papers following this develop the specific plans of two service groups for the continuation of psychological research on military problems of equipment design and operation[3] For these and similar programs, there may be merit in listing some observations on factors which made for success in wartime projects It will be desirable, in this regard, to distinguish between what will be called specific research, dealing with very specific instruments or devices with a singular use or purpose, and more general research, dealing with devices common to many work situations or purposes

First it should be noted that there was considerable economy and increased efficiency in specific research when all phases of that research were carried out by the same group It is readily seen that the problem of equipment operation and the problem of equipment design should be handled by a single group because they are really but a single problem Both are aspects of the task of fitting a job to a man To deal effectively with either, a psychologist needs the same background knowledge and the same research skills, and must use the same criteria But it is also true that when a particular research group has concluded an analysis of operating problems it has thereby acquired the best possible background for undertaking the supervision of the training program involving the same equipment This suggests the value of unified group work in the study, development, and introduction of new devices

Another fact which war research work made readily apparent was that when specific research is required, it is imperative for the psychologist to study in detail the basic engineering and functional characteristics of the device It is also necessary that he become completely familiar with any current doctrine which might apply to the operation or use of the gear Thus, the psychologist who becomes engaged in research on operating procedures and equipment design for particular devices ought to be about fifty per cent engineer He must be willing to dig into gadgets and learn what makes them tick Then he can talk intelligently to production engineers and ordnance personnel about the equipment and, what is more, he can plan his research and state his results in terms of the functional characteristics of the equipment

Because of the need for equipment analysis and because of the varied types of research problems which arise, a research staff usually functions more efficiently if it combines a wide range of talents and if its personnel represents a cross section of a number of scientific fields One reason why laboratory trained psychologists frequently did so well in handling war problems was that their training had been well distributed in other scientific fields By the same token, a balanced staff for any permanent research organization is implied

War research also showed the importance of the criterion of operator acceptance of new equipment or procedures No matter how satisfactory a design or procedure may seem as evaluated by other criteria, it has little value if service personnel reject it Reasons for rejecting a recommended device or method are many feelings of personal discomfort, biases established through previous training, rumored ill effects of operation or use It is well, therefore, to introduce tests of operator acceptance early in any equipment program In fact, developmental work in cooperation with small groups of operating personnel is highly desirable The end product, ready for field trials, then already bears a tentative stamp of user approval

In conclusion, it should be pointed out that wartime research groups bequeath to permanent research units a number of or

[3] One article is by F V Taylor and the other is by P M Fitts The former is on the Naval Research Laboratory and the latter concerns the A A F

ganizational problems in their applied work That these have been anticipated as continuing problems will become clear in the articles by Taylor and Fitts which fol low One of these problems is that of level of validation Should validation be made on the basis of full scale field trials, sim plified field trials, performance in simu lated action situations or performance in isolated tests of unit operations? This prob lem was ever present in the work of psy chologists during the years of the war and the solutions adopted were often expedients which will not suffice in more rigorous research programs Level of validation remains a matter of concern in every experiment on equipment design

Then, too, there is the need of establish ing more satisfactory ways of analyzing group performance and of setting up criteria and standards for measuring group performance Service units are for the most part teams rather than individuals This means that group coordination must be brought under more careful study Only when coordination or teamwork is ade quately measured can use be made of ap propriate group criteria in evaluation tests of new equipment or operating procedures

Decisions must also be made on the level of generality or specificity of research A program of specific research, seeking an swers to specific questions about particular pieces of equipment, is the more practical program when particular devices have been decided upon as the ones needed and when the time for research is limited If, on the other hand, one s only guide is a set of general plans for future equipment devel opment and there is time to explore fea tures which might be shared in common by many work situations then more general research can be planned Results can be stated in a way which will make them useful in the later design of elements or units in many and various kinds of devices Research which is less specific appeals to the psychologist because it fosters the de velopment of a more organized body of psychological knowledge But at the same time no small factor in his preference for general purpose research is that he may carry it out in the expectation that the results of his work will not be limited in application to national preparedness for war, but may see continuing use in the de sign of devices and materials for peaceful living

Men and Machines *

JACK W DUNLAP

The development of industrial psychol ogy in the past quarter of a century has been rapid but sporadic Critical demands during each of the two world wars ac centuated this development, but in differ ent ways From a psychologist s point of view, the outstanding need of World War I was to train masses of men quickly for different types of duty This need stimu lated the evolution of psychological meth ods of selection and placement The practical value of these procedures was appreciated by industry almost imme diately and in the post war period, many psychologists were invited to apply their newly aquired techniques to business and industrial problems In time, they added such other activities to their practice as progressive managers permitted Industrial psychology was, and is yet, predominantly a selection psychology

The distinguishing feature of World War II from a psychologist's standpoint was its highly technological character In this war, greater masses of men were trained for a greater variety of duties than in World War I, and the contributions of American psychologists in their selection and training have been well documented

* Reprinted from *Journal of Applied Psychology* Vol 31 No 6, December 1947

In the earlier part of the war, many different kinds of specialists sought to devise more efficient machines for seeking out the enemy and destroying him Many of these devices embodied highly desirable technical features which were impractical because the operation of these devices was too complex for the average man in the service With the largest American army and navy in history, there weren t enough sufficiently skilled men to operate or repair these devices Often, although suitable untrained personnel could be obtained, the time required for training was excessive Some radar repairmen, for example, needed more than fourteen months of training Even then, additional practical experience was considered desirable before they were ready for overseas service Under the exigencies of war, many of them acquired their practical experience overseas

Shortly after the beginning of the war, a small group of psychologists was asked to seek ways of adapting equipment to suit the operator instead of selecting men to fit equipment The results were so significant that the term human engineering began to creep into the vocabulary of *line officers* Before the war was ended, line officers not only listened to these psychologists, but exhibited a faith in their powers to a most flattering, if unwarranted, degree Such psychologists did more than articulate nicely the machine and its operator They were concerned also with the way one machine operator articulated with another close by, with the kind of information he needed to operate his machine, with how quickly he could get this information with a minimum of error, with the optimal conditions of operation, and other factors Starting with an interest in the interrelationship of men and machines, psychologists applied many of their professional techniques to problems which had only a secondary association with machines The approach was different from that of the time and motion engineer The time and motion engineer traditionally treats the machine as a constant and man as variable Optimal movements were explored which were most suitable to continued operation on a particular machine by a statistical or noncorporeal "average man" This group of engineers has contributed to the redesign of some equipment, but this activity has been marginal to the main trend of their work World War II, then, accelerated the beginnings of a new branch of industrial psychology It has been called bio mechanics, human engineering, bio technology, ' psychophysical systems research, and other names They all describe the planning of the machine, as E F DuBois has put it, from the man outward, considering the instruments and controls to be extensions of the man s nervous system (2, p 15)

The success of military and civilian psychologists in human engineering was acknowledged by design engineers and consulting groups working with the services Their continued interest in human engineering has serious implications for psychologists in these ways

1 Education of engineers along psychological lines

2 Training psychological personnel as human engineers to participate in the design of new equipment

3 Modification of the duties and capacities for service of the industrial psychologist

At the age of 40 years, 60 per cent of engineering graduates are in positions entailing administrative responsibility, according to Taylor & Boelter (9) This coincides with the aspirations of engineers as revealed by Karl Compton (1, p 71) On the walls of the national headquarters of the engineering societies in New York, there hangs the definition Engineering is the art of directing men and controlling the forces and materials of nature for the benefit of the human race' While some may feel that this definition is too broad and may cite examples of men not called engineers, who direct men and control (or try to control) the forces and materials of nature for the benefit of the human race, nevertheless, such men are really operating in the field of the engineer"

It is perhaps superfluous to point out that many engineers do a relatively fine job of administration even though they have had virtually no formal training in the "art of directing men' Engineers have received little formal training in this art

because there has been too little objectively verified material for them to learn The industrial strife of the past decade perhaps made them realize that the handling of *human* material sometimes decides the quality and efficiency of an engineering enterprise S A Lewisohn, in his excellent book Human Leadership in Industry, the Challenge of Tomorrow,' defines the problem clearly (7, p 48) He quotes Herbert Hoover as stating In these days of largely corporate proprietorship, the owners of mines are guided in their relations with labor by engineers occupying executive positions On them falls the responsibility in such matters, and the engineer becomes thus a buffer between labor and capital"

Lewisohn continues 'The question is What preparation have they had to act as such a buffer? A background limited to physics, chemistry, mathematics, mechanics, and other specific sciences does not equip a man to act as a buffer between labor and capital for their present responsibilities some training in psychological problems and the mental attitudes of men, some knowledge of modern sociological tendencies, some grasp of the incentives that make men act, some acquaintance with the purposes of trade unions and the art of collective bargaining, and some understanding of the technique of human engineering are indispensable"

In browsing through issues of *Mechanical Engineering* for the year 1946, I found no less than 22 articles expressing the engineer's concern with the problem of handling men It is a tribute to the engineer that he recognizes his shortcomings and would like to do something about them Articles on this subject were only slightly less common than those dealing with problems of atomic energy Some of these articles merely specified that a problem exists and something should be done about it Some writers recommended broadening the humanistic social base of engineering education Two articles contained rule of-thumb techniques of human management Two other writers stressed the need for management research and invited the engineer to apply his engineering training to such research

In none of these articles was there a report on the results of an experimental approach to the problem under the controlled conditions with which we are familiar in experimental psychology

I think this is significant because the engineer has been trained to solve problems by experimentation Idle discussion is foreign to his tradition These articles indicate that the administrative engineer *wants* to solve social problems and lacks the know how to solve them This deficiency is not to his discredit Sociologists, economists, labor relations specialists, and psychologists have puzzled over these problems for years and have made not much greater headway Clearly, this is a vast problem, which requires a co operative approach by all of these professions In my opinion, the psychologist can make a particularly strong contribution to their solution, not because he possesses any peculiar set of facts, but by the application of principles and techniques developed or tested during World War II The psychologist should have the opportunity to apply his knowledge to this industrial problem on a sufficiently large scale to make his research significant

Broadening the training of the engineer in social humanistic subjects is a current trend Rensselaer Polytechnic Institute has increased by one third the time given to psychology and other humanistic studies Newark College of Engineering has reorganized its curriculum to place greater emphasis on human problems The Department of Engineering, University of California, Los Angeles, California, is expanding its curriculum to include two 'bio technical courses The first of these, *The Dynamics of Human Function and Behavior,* deals with practical psychology, and with the physical structure, thermo dynamics, and machinery of the body The second course, *The Influence of Environment on Man,* delineates his interaction with the atmospheric, thermal, bacterial, radiational, and chemical aspects of the environment, and includes a study of socioeconomics (9) The psychologist, at this time, can make the soundest contribution to the content of such courses by exploiting the field of biomechanics

The engineer can no longer be satisfied with "fatal limits in extending his control over the environment of man Dr DuBois

outlines the idea in this way Perhaps we are wrong in trying to place a sharp limit on the factors of safety and devote so much attention to the fatal level It may be useful to know the fatal level, but the engineer should concentrate on the levels when men first become inefficient If a certain machine, such as an airplane, is under the control of a man, its efficiency corresponds with that of the man If in a sharp turn a pilot loses his vision, both man and machine are blind If the pilot at an altitude of 38 000 feet develops intolerable bends, he and the plane must descend to a much lower altitude At extremely high or low temperatures, there is a marked loss of mental capacity as well as loss of muscular power and control " (3, p 627) The psychological literature contains many verified data on the effect of environmental change on performance Such information should be helpful to an engineer s training

Psychologists at universities with engineering schools may anticipate a demand for courses in psychology suitable for engineers Such courses would (1) make available to the engineer pertinent psychological data, (2) impress on the engineer the importance of considering limitations of the operator in designing equipment, (3) describe to the engineer the kind of research techniques that have been useful to the psychologist, and (4) form a basis for mutual understanding, respect, and intelligent co operation on problems peculiar to each group

Some psychologists have expressed concern about the possibility of competition between the engineer and the industrial psychologist I believe this concern is groundless Engineering schools have enough trouble in training a good engineer in 5 college years, 7 or 8 years is not too long a time for training a good psychologist Ultimately, the engineer and the psychologist will have to work together, at least in the bio technical field The technical scope is too vast for one type of professional person It would be wiser for the psychologist and the engineer to acquire a mutual appreciation of their technical skills and understand their own limitations

To summarize what I have said, engineers have educed an acute interest in the techniques of dealing with men Through their experience in World War II many engineers have been impressed by the psychologist s contribution to machine design In fact, during committee deliberations in 1945–46, regarding the establishment of a Science Research Foundation, engineers supported psychology as a science to be included with the physical sciences Psychology must become a part of the curriculum of the engineer, and psychologists will be expected to co operate in preparing or giving suitable courses

Let us consider now the opportunities for the psychologist trained in human engineering techniques A great amount of work on equipment design, carried on by military agencies, has been summarized by Kappauf (5), Taylor (8), Fitts (4), and Kelly (6) The opportunities for employment by these service agencies require no further elaboration

We may be about to experience the greatest change in style of life since the Industrial Revolution The full utilization of atomic energy for peacetime pursuit will bring us close to supersonic speed, interstellar space and exploitation of the mineral wealth in the arctic zones Man will work and live in strange worlds, thus new and unusual problems will confront the psychologist For example, consider a simple problem in vision in thrust craft operating at a speed of 2 000 miles per hour, or about 3 000 feet per second Visual stimuli initiate nerve impulses which reach the optic cortex in about 0 05 second If an observer in such a craft were to look at an object directly abeam, he would have flown 150 feet beyond it before his brain registered the stimulus Such an observer would probably feel a bond of sympathy with the mythical bird who flies backwards because he wants to see where he has been Seriously, the design of control devices by means of which an operator can govern such craft offers a real challenge to psychologists Because of the high altitudes at which such craft may operate, the entire field of visual science will have to be restudied, for there is reason to believe that the upper sky is always dark This, in turn, poses problems in high altitude navigation and camouflage

As natural resources are consumed, man will be tempted to find and remove the wealth of the arctic regions. This, in turn, will create an extensive series of physiological, sociological, and psychological problems with regard to the action and interaction of human organisms under extreme environmental conditions. These and hundreds of other similar problems may seem fantastic, but I am convinced that most of us will live to see them treated as routine.

Fascinating as it is to speculate about our future, there are many industrial problems which require the skill of psychologists right now. There are few industries which appreciate the usefulness of the psychologist as much as the aircraft industry. Yet there is recurring evidence of the need for more of the psychologist's services. One day, not so long ago, I visited a plant in which jet fighter planes are being built. As I stood near the end of a runway, I saw a towing crew working on a crumpled automobile. There were other crumpled cars near it, and my curiosity was aroused when I counted 12 badly damaged cars in that parking lot. I learned that the day before one of the jet planes was rolling along the runway just after landing and that the brakes when they were applied by the pilot would not hold. It must have been embarrassing to him, for many of those planes land at 120 miles per hour or more. Why the brakes failed does not matter to us, so let us consider what the pilot could have done. He could have ground looped, that is, touched one wing to the ground and taken his chances of survival but there were workmen along the edge of the strip, and they might have been injured. It appeared that the wisest thing to do was to retract the wheels and slow down by skidding along the plane's fuselage. In this particular plane, the "wheel up" lever is behind and to the left of the pilot's seat, so that it is operated after a blind search by "feel." To make certain that it is not inadvertently tripped by the elbow during flight, a safety cover has to be raised before the lever can be operated. The pilot tried to manipulate the switch, failed, and tried again, but by this time, he had rolled off the runway and into the parking lot. The plane smashed into car after car and finally turned over supported by a car at each wing tip. Fortunately, the pilot escaped with minor injuries and climbed out of the plane before rescuers arrived. Two things seemed clear: there should be emergency brakes and if "wheels up" is to be considered as an emergency landing procedure, this control should be placed where the pilot can get at it quickly. This control is in an inconvenient and dangerous position, and yet it has to be used twice in every flight. After the accident, it was obvious that the control lever was improperly placed. This example emphasized the need for examination of equipment in the design stage from the viewpoint of human abilities and limitations. During the war many psychologists helped engineers by examining new equipment in an attempt to prevent just such occurrences.

The importance of vision in industrial plants has been dramatized by dispensers of safety goggles by means of such instruments as the "Sight Screener," and the "Orthorater." These are not the only ways in which the psychological principles of vision can be employed. The law of contrast has been used widely by safety engineers in painting dangerous areas and machines. The same law can be applied to the function of a machine. In the pharmaceutical business "paddles" are employed for counting tablets. A paddle is an unpainted piece of aluminum or wood in which a predetermined number of holes have been bored part of the way through. The holes are slightly larger than the tablets to be counted. The operator shoves the paddle into a container of tablets to pick up a load, and then shakes it with a wrist motion until each hole contains a tablet. This mechanical method of counting is simple and efficient, but it is not foolproof. Counting errors are made, and these appear to be the fault of the operators. An operator occasionally fails to perceive that one of the holes does not contain a tablet and pours a short count into the container. This perceptual error is due, in part, to the fact that the color of the tablets and of the paddle do not contrast sufficiently. Inspection precision was improved simply

by placing a spot of contrasting color, somewhat smaller than the diameter of the tablets, in the bottom of each hole

Motor habit patterns have been investigated by psychologists for decades A great deal is known about establishing such patterns quickly and efficiently, but very little work has been done on the extinction of motor patterns There is a real and practical need for additional experimental work on the extinction of habit patterns, for the results can be applied directly to industrial and military situations In cases of emergency, surprise, or fatigue, an individual tends to revert to earlier inappropriate motor patterns This fact is not well known, nor is it usually considered by designers of equipment, possibly because the problem has not been recognized by engineers Today, a driver can operate any American passenger automobile without being confused by the gear shift, but only a short time ago there was the Buick shift the Dodge shift, the standard shift, and the Ford planetary drive The lack of standardization creates an accident hazard based on motor habit patterns That it is not a problem of the past was shown by the recently publicized accident of the Royal Dutch Airlines, in which the pilot pulled what he thought was the flap retractor lever but actually was the landing gear lever The result was a serious accident in which several persons lost their lives and the plane was almost completely destroyed The pilot knew where the flap retractor lever was, he had been checked out in the plane, he was familiar with it, but he reverted to an old motor habit established in another type of plane This is not an isolated error, it has occurred again and again, both in civilian and military aircraft Pilots flying aircraft with which they are relatively unfamiliar have switched to empty gas tanks, cut the ignition of good motors when they intended to cut bad motors, or have feathered the wrong propellers when engines failed Such errors can be obviated by standardizing the cockpit, and by the application of psychological data to the principles of design as they relate to the operator Elimination of earlier motor patterns is not so simple, but the need for work on the general problem is extremely urgent

The entire transportation industry is filled with problems concerning the individual and the machine, not only from the standpoint of operation and safety, but also that of comfort of the passenger Railroads have been severely criticized for their lack of consideration for the passengers comfort True, railroad companies have promised wonder trains for the post war era, but for the most part, these are still on the drawing boards, or at least only pilot models are available The problems of lighting, noise, and heating cannot be solved by current engineering methods alone, but must be solved, in part, in terms of the human factor Reducing the noise level in a railway coach to a given number of decibels is not enough Noise should be expressed in such terms as 'Is the noise level sufficiently low so that passengers sitting in adjacent seats can converse in low normal tones? I could go on discussing this specific problem, but my purpose is merely to stimulate psychologists to think in terms of applying their knowledge to transportation problems When they do, they will make a real contribution to the travelling public Buses and street cars and that mighty denizen of the open road, the truck, all need careful scrutiny Some problems are common to all of these, but each has large groups of problems peculiar to itself For example, consider cross country trucks, with their problems of relief drivers, fatigue, road strain, cab design, seat design, and bunks for relief drivers, to mention only a few of the problems involving the operator which need to be studied

The radio manufacturing industry can profitably use the services of design psychologists One company questioned the advisability of automatic volume controls in receivers The first question their engineers proposed to the psychologist was, Do radio listeners need and want automatic volume control? It was found that listeners desired this feature, so the next question was, 'How sensitive must the control unit be? Basically the problem was to determine the point at which individuals would adjust volume control, regardless of

the level at which they initially set the volume Experiments proved that the volume could vary as much as 4 decibels before the listener would readjust the volume The psychologist thus was able to provide the engineer with definite limits of sensitivity for designing the automatic volume control unit

The psychologist has given little thought to the architecture of homes, stores, office buildings, theatres, and other places of public assembly One need only examine any major occupation, such as transportation, building, manufacturing in all its forms, or even farming, to observe innumerable problems involving the interaction of men and machines

Within a short time, therefore, I predict that more psychologists will be sought for the solution of bio mechanical problems on a full time basis in the automobile, radio, home appliance, transportation, and other industries Their principal contribution will be to provide research in equipment design

Equipment design is but one aspect of bio mechanics You will recall, that during World War II, the human engineer started with a narrow man machine relationship and found it necessary to extend his study The man machine relationship was a point of departure The influence of World War II on industrial psychology probably will be even greater than the influence of World War I It is my opinion that industry is aware of the many problems which can be solved by human engineering Managers will seek industrial psychologists for this type of service The industrial psychologist will have to add to his technique of selection and training the techniques of adapting machines to man

In the broadest sense, the industrial psychologist is concerned with the development of a more productive society Greater attention must be given to the articulation of man and machine to achieve greater industrial production, and so, the industrial psychologist must be prepared to delve into the remotest phases of this particular problem It is often difficult to determine where bio mechanics ends and other aspects of industrial psychology begin For example, deliberate slow downs may be caused by the application of time and motion methods, or by the installation of more efficient equipment The attitude and morale of a worker, then, are important A tire manufacturer installed some new equipment Motion studies in other plants had shown the average worker could process 24 tires per hour on this equipment After training, he found his workers were producing only 17 tires an hour and later the average dropped to 13 A review of the work method indicated minor adjustments, including additional pay which finally brought production up to 15 tires per hour The plant never reached 24 per hour because of a deliberate worker slow down to prevent the possible lay off of some workers Clearly, that manager failed to recognize the psychological effect of new equipment and new methods on the social environment of the workers If these problems, with their social overtones of the relationship between men and men (and I say men and men advisedly, not labor and management), had been recognized as were the relations between men and machines, it would have been possible to develop a transition program for the workers, based on psychological principles Such a situation can be alleviated after it occurs, but only at considerable expense in terms of human emotions and production losses, which could have been avoided by proper preparation

A great deal is written about industrial safety programs and the use of automatic safety devices There is no question about the need for such devices, but the interaction of men and machines is not perfect Recently I was advised of an amusing incident related to safety devices, and I say it was amusing because no one was seriously injured before the source of difficulty was identified In a plant which prides itself on its safety program, women operate a heavy stamping machine When the power pedal is activated, a vertical shield drops in front of the mechanism and an upright bar moves horizontally across the face of the machine to shove aside the hand of the careless operator These safety devices were painted in conspicuous colors, for the plant makes wide use of color in its safety program One day, shortly after

the equipment was installed, a girl operator suddenly fainted Thereafter, scarcely a day went by when one or more girls did not give up and quietly slide to the floor No amount of physical examination or study of medical history gave any clue to the phenomenon Finally, a psychologist who had worked for the plant was called in and succeeded in finding the cause after a little study of the working conditions The monotony of the task, and the movement of the colored safety mechanism set up a condition required for hypnosis Once the difficulty had been identified, the solution was easy I might say that the psychologist s professional stature was not reduced

Safety devices for emergency control are not adequate unless they are immediately and readily available to the operator Often such controls are added as afterthoughts, and are located because of engineering convenience rather than the convenient use of the operator During the war a series of accidents occurred in Navy planes, in which the pilot s head came in violent contact with the gun sight during landings aboard carriers To eliminate such accidents, a shoulder harness was devised to hold the pilot so that his head could not come in contact with the gun sight Unfortunately for the pilot, it was necessary for him to lean forward and down into the cockpit to adjust a lever during landing It was impossible for him to do this while he was in the harness, so he released the harness and took a chance with the gun sight The shoulder harness was perfect for the job it was designed to do, but the engineer forgot the man and what the man had to do

Recently, I heard of an interesting problem in ' selection ' An industrial organization had an entire contract cancelled because five items in a lot of 100 gross were defective These items were worth only a fraction of a cent each, but the full contract ran to thousands of dollars The management decided, What we need is a good selection system for inspectors ' Therefore, a psychologist was called in to establish a selection system The first question to be answered was Selection for what? So, a thorough study had to be made to ascertain why defective items slipped by the inspectors The items passed before the inspectors on a conveyor at the rate of 200 per minute If an item was defective, the inspector removed it from the belt and placed it in a container at her side When the container became heaped up, she would bend down and level the pile of material, and her eyes were away from the belt for from one to ten seconds Many of the younger inspectors were likely to watch the young foreman as he moved about the department Although selection might help, particularly if elderly, unattractive men were the object of selection, the fundamental problem was to determine the attention distractors and to find means for reducing or eliminating such distractions Some remedies are immediately obvious such as rearrangement of the work stations with regard to the disposal of defective items, training in the use of peripheral vision for handling defective items, and so on

The principles underlying bio mechanics and the techniques employed by workers in this field can be applied to the practical problem of conserving raw materials and increasing production An ever present, costly problem confronting a manufacturer is that of quality control, which can be restated as a problem in the control of the variability of the product This variability is a function of the raw materials, the machines on which the product is fabricated and the workers who fabricate the material It is a common practice to examine the variability of the raw product, and in many plants to exercise some control over the machines Only rarely is the human factor seriously considered as a contributing cause, and this is particularly true if the machines are automatic or semi-automatic The first problem is to identify the contribution of each of these factors to the total variability The part of the total variability which is a function of the inter action between the three major sources of variation should then be determined

During the past months several members of the Division of Bio Mechanics have worked in greige mills producing nylon hosiery The general objective was to in-

crease production and decrease consumption of raw materials A series of carefully controlled investigations was instituted to determine what part of the total variation could be attributed to the raw nylon yarn, the machines, and to the operators They found that the largest sources of error were contributed by the machines and the operators Over all plant losses as high as 8 per cent were caused by these two factors in addition to the normal wastage factor, and often neither their source nor nature was suspected Further studies revealed more than a hundred ways in which well trained knitters unwittingly can cause their stockings to vary For young or inadequately trained knitters, the variation was exaggerated

Once the relative contributions of these factors had been determined the problem resolved itself into preparing a remedial program for the operators and the equipment The first step was identifying the various operations which the knitter could perform and determining their effect Once this was done, a sound, practical, and easily administered retraining program could be devised This approach to the control of quality is a completely general method and can be applied to many kinds of productivity other than the manufacture of hosiery

These are but a few examples of the variety of problems with which the human engineer may be confronted The overlap in approach to industrial problems by the human engineer and the industrial psychologist is not the only reason for a modification of the practices of the industrial psychologist

When managers require psychological services, they usually do not know enough about the specialties of the industrial psychologist to seek one type of person to handle a problem of selection, another a problem of human engineering, and a third a problem of training Managers expect a psychological agency to handle most of the problems they regard as psychological Correspondingly, it is reasonable to anticipate that managers will expect the industrial psychologist on their staff to become as competent in this new activity as he is in selection and training In the current operation of the Bio Mechanics Division of the Psychological Corporation, we have found it necessary, at times, to go into problems of training, of selection, of location of work stations, of the study and modification of other physical aspects of the environment such as temperature, ventilation, noise level and vibration, lighting and color We found it necessary to concern ourselves with the breakdown of a task for more equitable and more efficient distribution of effort among men and women working on machines We were also concerned with the techniques of operating machines, of worker morale of labor management relations and time and motion studies Indeed, we are expected to tackle almost any problem of production in which our understanding of the functions and limitations of human beings may lead to a solution

All of us have seen machines grow in complication until they have become a psychological burden to the average operator It has been extremely important to collect scientific information about the interrelationships between men and machines with the objective of enhancing the efficiency of the machine and the comfort of the operator Perhaps a more subtle but no less complicated alteration has been made in the social structure of our agencies of production It is possible that most of the strife between labor and management, so costly to society as a whole, is a symptom of a too complicated or an inadequate social milieu in the industrial plant Perhaps the mutual isolation of labor and management, of labor's insufficient feeling of participation in the production of the finished item, of an insufficient feeling of prestige and worth, to mention but a few possibilities, are determining factors Here, too, we need to collect scientific data to determine whether the social structure of the factory has impinged upon a human limitation, and if so, where, we should not limit ourselves only to sensori motor limitations We need precise information about the differences in the morale of workers at different plants The psychologist has a special advantage in approaching such problems by virtue of the rich techniques which are part of his tradition

Today, those of us who are working in bio mechanics have drifted into the field on currents activated by a wide variety of interests None of us received systematic training for this work and only chance or temperament has fitted us for it We should now consider more systematic training for this aspect of applied psychology

As a starting point for a discussion of a course of training, I believe that most of the basic courses are now being given by many departments The *names* of such courses are not sufficient, it is the emphasis given in such courses that is critical For example, there are numerous courses in introductory statistics, such as statistics in agriculture, in education, in psychology, in economics, or in biology, and all of these basic courses contain essentially the same fundamental logic of analysis and elementary formulae Yet, a student of one of these courses finds it difficult to apply his knowledge to the problems in another field This difficulty depends on the direction of the course, which usually is determined by the experience of the instructor and the examples used for demonstration Thus, we do not necessarily need new and different courses but new examples with a new emphasis on the applications of the skill and techniques developed by the courses The present courses of business and industrial psychology, physiological psychology, experimental psychology with a particular emphasis on practical problems, educational psychology with emphasis on the problem of learning and retraining, statistics, tests and measurements, interests and attitudes, and a course in the design of experiments all provide basic training for work in bio mechanics Such courses are not sufficient, however, unless they are deliberately aimed toward the application of psychology in human engineering Certainly the field is sufficiently well defined to work out a year's graduate course in bio mechanics Ultimately, there is no substitute for practical experience We, in the Bio Mechanics Division of the Psychological Corporation, have been happy to contribute to a partial solution of the problem by providing interneships for properly qualified graduate students In addition until there is enough literature on the subject, it would be desirable for some university staff members teaching graduate courses in statistical, industrial, physiological or experimental psychology to spend a year with some agency involved in a variety of human engineering problems

In summary, the development of psychological activities during the latter part of World War II has implications for the expansion of psychological training in the education of the engineer, the development of a new specialty which involves the application of psychological data and principles to equipment design and operation, and the future development of industrial psychology

References

1 Clyne, R W, *Engineering Opportunities* New York Appleton Century 1940 Foreword, K T Compton

2 DuBois E F The Anatomy and Physiology of the Airplane Cockpit *Aeronautical Engineering Review* April 1945, Vol 4, 15

3 DuBois, E F, Limits of Factors of Safety in the Human Body *Mechanical Engineering* 1946 625–627

4 Fitts, P M, Psychological Research on Equipment Design in the AAF *American Psychologist* 1947 Vol 2, 93–98

5 Kappauf, W E, History of Psychological Studies of the Design and Operation of Equipment *American Psychologist* 1942 Vol 2, 83–86

6 Kelly, G A (Ed), New Methods in Applied Psychology *Proc Maryland Conference on Military Psychology* University of Maryland, College Park, Maryland, 1947, 137–170

7 Lewisohn, S A, *Human Leadership in Industry the Challenge of Tomorrow* New York Harper & Bros 1945

8 Taylor, F V, "Psychology at the Naval Research Laboratory,' *American Psychologist* 1947 Vol 2 87–92

9 Taylor C L and Boelter, L M K, Biotechnology A New Fundamental in the Training of Engineers," *Science* 1947, Vol 105, 217–219

Chapter VIII

DESIGN OF DISPLAYS

A display, according to Fitts, is any device that can be used to present information to individuals by visual, auditory, tactual, or other exteroceptive channels The design of an effective display is the concern of a psychologist as well as the engineer, since the information transmitted by the display varies in accuracy according to such variables as size, shape, spacing, pattern discrimination, and so forth

Continued engineering development of such machines as the airplane and automobile creates problems of display design Instruments must be designed to indicate and transmit information and those that are capable of being read most rapidly and most accurately help the operator of the machine to do the best job Research in display design of airplane instruments has gone far beyond that for automobiles Although the instrument panel in automobiles may not be as important as in airplanes, it is, nevertheless, important enough Variation in shape, size, lighting, location, and scale markings are a few of the items that have been designed with or without considering the driver and his effectiveness It appears that rigid engineering tests of motor performance and consumer preference tests of aesthetics of design are of greater concern to automobile manufacturers than the effectiveness of displays

The studies selected for this chapter illustrate some of the problems in connection with display design It is to be noted that only one is an industrial study The remaining four originate in the field of "military psychology" and are concerned with displays in aviation Most of the work in engineering psychology stems from this source As its effectiveness is demonstrated and as industry becomes aware of its existence then one can predict that research more directly related to industry will be undertaken After all, it is difficult to conceive what the present status of psychological testing in industry might be had it not been for the impetus it received during World War I

Grether's study illustrates the value of experimentation in the design of quantitative displays In a careful laboratory experiment, he demonstrated how such work can lead to very practical results By testing the reactions of two groups of subjects to nine altimeter designs he determined that differences in accuracy of readings occur The conventional altimeter was found to be difficult to read, and a more suitable instrument was recommended Grether also determined that speed and accuracy of instrument reading are positively related—instruments that can be read more quickly can also be read more accurately Instrument-reading difficulties are not obliterated by experience since college men without experience obtain results similar to those with considerable Air Force experience

Radar is an amazing development but its effectiveness is limited by the accuracy of the humans who do the readings Ford finds that errors in readings are reduced by introducing finer scaling but that two types of systematic errors then influence the reading of scales The confusion error is sufficient not to warrant the use of finer scaling in the situation studied This study is valuable because it indicates that an improvement in one part of a situation does not always mean that all other factors have not changed For example, one might improve brakes in

automobiles but then introduce a new difficulty of having the passengers propelled violently forward whenever the brakes are applied The Grether and Williams study has been included for its experimental design as well as its findings concerning errors of interpolation in relation to the variables studied

Lawshe and Tiffin checked on the accuracy of using micrometers and calipers Their findings clearly indicate that the accuracy of precision instrument usage is much less than expected Can it be that the design or scale of such instruments is mechanically unequal to human abilities and limitations? If so, then a series of experiments paralleling those of Ford and Grether reported here might be appropriately devised to improve the accuracy of such readings It may well be that such work would suggest changes in equipment design, in addition to further training as recommended by Lawshe and Tiffin

The displays studied in the preceding articles were visual It is recognized that displays may present information to other senses, and the article by Forbes indicates the practicality of using auditory displays The Flybar or *fly*ing *b*y *a*uditory *r*eference is the name for an auditory display suggested as a substitute for visual displays especially when the eyes are subjected to an overload A three in one auditory signal was devised and the article indicates that subjects learned to "fly" Link Trainers with only auditory cues for guidance

Instrument Reading I The Design of Long-Scale Indicators for Speed and Accuracy of Quantitative Readings *

WALTER F GRETHER

The data presented in this paper have been previously reported in Memorandum Reports No TSEAA 694 14 and MCREXD 694 14A of the Aero Medical Laboratory, Engineering Division, of the USAF Air Material Command

Quite a number of instruments used in aviation and elsewhere must be read with precision greater than can be provided by one revolution of a pointer on a circular dial of conventional size There is considerable accumulated evidence that, except for the direct reading counter, most of the devices that have been used to increase effective scale length result in instruments that are very difficult to read In a previous study by the author (2) on the design of clock dials, it was found that as common an instrument as a clock is quite difficult to read Even the best clock designs required approximately 5 seconds (including recording time) for readings in hours and minutes by Air Force pilots Even with this time spent on each reading, about 7 per cent of the readings on the better clocks were in error

Aside from such laboratory data there is considerable evidence of instrument reading difficulties in the practical situations where these instruments are used In a study of actual errors made by pilots in reading aircraft instruments carried out by Fitts and Jones (1), multiple pointer or long scale instruments provided the greatest number of serious cases of instrument misreading The instrument reported as being misread most frequently was the altimeter In the typical report the altimeter was read too high by a complete

* Reprinted from *Journal of Applied Psychology* Vol 33, No 4, August 1949

revolution of the most sensitive pointer, that is by 1000 feet A tachometer designed with a rotating sub dial to indicate RPM in thousands was likewise read too high by 1000 RPM Numerous fatal and non fatal accidents have been attributed directly to such instrument reading errors, and without doubt many of the unexplained crashes resulted from similar human failures

The major purpose of the present investigation was to make a direct comparison in terms of speed and accuracy of quantitative readings of several of the possible methods of obtaining increased scale length on instruments The experiment also had a secondary but more specific and practical purpose of finding improved methods of indicating altitude in aircraft For this reason all of the instruments were designed to read altitude in feet and all readings were made in feet as units

It is emphasized that the evaluation of the different indicator designs in this investigation was with respect to the speed and accuracy of quantitative readings Actually this is only one of several criteria which most instruments should be required to satisfy It has been pointed out by the author (3) that in aviation in particular there would appear to be at least three major ways in which instruments may be read, depending upon the purpose of the reader These three types of reading may be categorized as follows a Check reading—for assurance of a null, normal, or desired indication, b Qualitative reading—for the direction and approximate magnitude of a deviation from a null, normal, or desired indication, and c Quantitative reading—for the numerical value of an indication

The above categories of instrument reading have considerable utility as criteria against which to evaluate different instrument designs It is usually possible from a knowledge of the situation in which an instrument is to be used to decide the reading purposes or criteria which it is most necessary to satisfy The criteria against which an instrument is to be evaluated then provide operational definitions of the experimental measurements to be made As mentioned earlier, the experimental indicators in this investigation were evaluated only with respect to the third criterion quantitative reading In this study, furthermore, there was no concern with small errors of interpolation, only with larger errors resulting from assignment of incorrect values to graduation marks

Experimental Procedure

Nine experimental altitude indicator designs were used in this investigation These are shown along with some of the results in Figure 31 1 The first of these indicators, design A, is a simulation of the altimeter almost universally used in military and larger commercial aircraft On this instrument the longest pointer gives readings in hundreds of feet the broad pointer is read on the same scale in thousands of feet, and the small pointer is read on the same scale in ten thousands of feet Altimeter designs B and C also simulate existing but not commonly used types

Altimeter design D uses a single pointer to indicate altitude in hundreds of feet This pointer makes one revolution for each 1000 feet change in altitude and the multiples of 1000 feet are indicated on a simulated direct reading counter This counter has two drums, one for 1000 foot and the other for 10,000 foot increments It is assumed that the motion of these drums would be intermittent and that single whole numbers would always be showing

In design E, also, only one pointer is used, but two dials rotating behind a window indicate the multiples of 1000 feet In this design the motion of the dials showing through the window is assumed to be continuous rather than intermittent, thus permitting more than one number (or half numbers) to appear

Design F indicates altitude in quite a different manner from the other instruments In this display the pointer is assumed to make only one revolution to cover the entire altitude range The range being covered is indicated in the window as 0–1000 feet, 0–10 000 feet, or 0–100,000 feet The meaning of the numerals on the dial graduations is, therefore, determined by the range indicated in the window

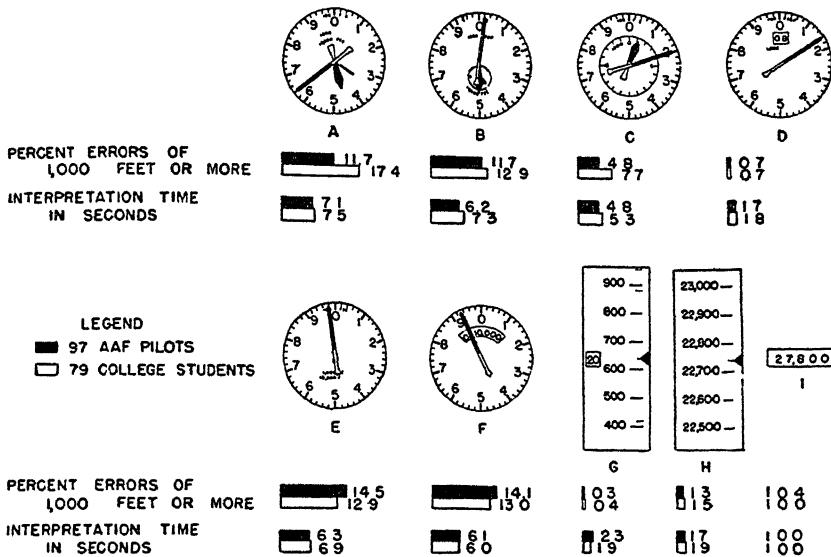



FIGURE 31 1 *Speed and accuracy in reading altitude from different types of instruments*

This indicator is similar in principle to a radio altimeter now in use It is obvious that the precision of indication on such an instrument decreases as the range being covered increases

Altimeter designs G and H are similar in that they simulate a scale moving vertically behind a window An instrument following design G could use either an endless tape or drum to present the moving scale, with a counter to indicate multiples of 1000 feet An instrument using design H would require a very long tape with a scale covering the desired altitude range

The last experimental design, I, simulates a simple direct reading counter without any moving pointer or scale For reasons pointed out later in the discussion of results, such an indicator would probably be unsatisfactory for the pilot, but might be suitable for other aircrew members such as the navigator One of the major reasons for including it in this study was to get an approximate measure of the time required to copy a series of numbers representing an altitude reading, it being assumed that no interpretation time would be involved in reading altitude from this type of indicator

For each of the altimeter designs used in this experiment a test booklet was prepared The cover (page 1) of each booklet presented the experimental subject with detailed instructions for reading the dial design in that booklet, and a sample dial for the subject to read On the two inside pages 2 and 3, the dial design was reproduced with 12 different settings Under each picture was a space for writing in the reading [1]

Special precautions were taken in the preparation of the drawings and choice of altitude settings to be used in the various test booklets to prevent biasing the results for or against any of the indicator designs The circular dials were 2¼ inches in diameter From this other dimensions can be estimated from Figure 31 1 All essential numerals and graduation marks were sufficiently large and distinct to be easily legible Except for the inner dials on designs B and C all scales were alike in hav

[1] The large number of drawings needed for the nine test booklets were produced by Miss Mary Cowles of the Psychology Branch with the photographic assistance of Mr D M Penrose of the Laboratory Services Unit of the Aero Medical Laboratory

ing numerals at all 100 foot graduations with intervening marks at 20 foot intervals Other factors equalized were the number of settings above and below 10,000 feet, the number of sensitive pointer settings on 100 foot graduation marks, the number of sensitive pointer settings just preceding and just following the zero on the scale, and the number of sensitive pointer settings on the left and right halves of the dial Precautions were also taken to be sure that no essential information was hidden by any of the hands, and that the interrelationships between pointer positions were correct For indicator design F some of the settings were midway between graduation marks For the remaining designs the sensitive pointer (or reference mark) was always on a graduation mark Thus, no interpolation was required to obtain correct readings

The altimeter reading test was taken by 97 USAF pilots in the Instrument School at Barksdale Field, Louisiana, and 79 college men (without aircrew experience) at Denison University, Granville Ohio In administering the test, the booklet for only one altimeter design was passed out at a time, and sufficient time was allowed for reading the instructions and working the sample item At a signal all subjects opened the booklet and worked until completing all items Each subject s completion time was recorded on his booklet Four sequences for administering the nine test booklets were used in order to counterbalance for learning effects An approximately equal number of subjects (in each of the two subject groups) took the test in each sequence

The two subject groups of dissimilar experience were used in order to get some measure of the effect of experience on the ability to read the various dial designs All of the USAF pilots can be assumed to have spent several years flying with altimeter design A, and possibly some experience with designs B and C The college men can be assumed to have had little, if any, experience in reading altitude from any type of indicator In general intelligence and education the two groups were very similar

Results of Comparisons Among Indicator Designs

The data obtained in this investigation were analyzed to determine the frequency of errors and the time per instrument reading These results are shown in Table 31 1 which gives the per cent of total readings in error for the nine indicator designs [2] None of the errors included in this table resulted from inaccuracies in pointer interpolation since all settings of the sensitive pointers were on graduation marks (except for design F which had some settings midway between marks)

Data on speed of reading are also shown in Table 31 1 It will be recalled that the subjects wrote their answers in the test booklets and the time for completion was recorded in each instance The average time per reading could thus be computed from the total time and the total number of items, but this time included the time for recording as well as for reading or interpreting the instrument

A reproduction of each of the experimental indicator designs accompanied by graphic illustrations of the more significant findings is provided in Figure 31 1 The upper pair of bars under each indicator shows the per cent of errors equal to or exceeding 1000 feet for the two groups of subjects The lower pair of bars gives the computed interpretation time for each of the two groups of subjects An estimate of the time for interpretation only was obtained by subtracting from the average time for each design the average time for design I (the direct reading counter) The reading of altitude from design I involved the mere copying of the numbers shown, and hence was assumed to require no interpretation

Discussion of Results

Indicator designs A B and C The results of this investigation, as shown in Figure 31 1 and Table 31 1, show that design A, which simulates the conventional

[2] Altimeter reading errors during actual flight probably occur with much lower frequency than found in this study, since in flight the pilot can anticipate the approximate readings

TABLE 31 1

Altıtude Indıcator Desıgn	USAF Pılots N = 97		College Men N = 79	
	Per Cent Errors (a)	Seconds per Readıng * (b)	Per Cent Errors (c)	Seconds per Readıng* (d)
A	15 9	9 6	20 8	9 8
B	15 0	8 6†	17 9	9 6
C	8 3†	7 3†	11 4†	7 6†
D	3 5†	4 2†	2 1†	4 1†
E	17 3	8 8	15 3†	9 2
F	24 1	8 7†	21 0	8 3†
G	2 1†	4 8†	3 0†	4 2†
H	2 5†	4 2†	4 5†	4 2†
I	0 6†	2 5†	0 3†	2 3†

	N	r	Confidence level
Correlatıon between speed and accuracy for dıfferent desıgns			
For pılots (columns a and b)	9 desıgns	91	1%
For college students (columns c and d)	9 desıgns	95	1%
Correlatıon between pılots and college students on dıfferent desıgns			
Per cent errors (columns a and c)	9 desıgns	95	1%
Seconds per readıng (columns b and d)	9 desıgns	99	1%
Correlatıon between speed and accuracy of ındıvıduals on all desıgns			
Pılots	97 pılots	38	1%
College students	79 college students	44	1%

* Readıng tıme ıncluded tıme for subȷect hımself to record answer

† Indıcates statıstıcal sıgnıficance (at one per cent level of confidence) of superıorıty over conventıonal altımeter (desıgn A)

altımeter, ıs a very dıfficult ınstrument from whıch to obtaın quantıtatıve readıngs as re quıred ın thıs study Even the pılots, all of whom had spent several years flyıng wıth thıs ınstrument, spent more tıme per read ıng on thıs ındıcator than on any of the other desıgns studıed Only 1 of the remaın ıng 8 ındıcators, desıgn F, resulted ın a hıgher proportıon of errors It must be con cluded that ıt ıs a very dıfficult task to combıne ınto a sıngle numerıcal value the readıngs of 3 poınters ındıcatıng on a sıngle scale, as requıred ın readıng the conven tıonal altımeter Desıgns B and C appar ently were slıghtly easıer to read than desıgn A

Indıcator desıgn D [3] Thıs ındıcator uses only one poınter, wıth the 1000 foot and 10,000 foot ındıcatıons provıded by a counter Such a combınatıon proved to be very easy to read For USAF pılots the per cent of total errors was very low, 3 5 per cent, and only 1 7 sec was requıred for ınterpretatıon (as contrasted wıth 15 9 per cent and 7 1 sec for the conventıonal

[3] On the basıs of thıs study ındıcator desıgn D, combınıng a sensıtıve poınter wıth a dırect readıng counter, was recom- mended as a replacement for the exıstıng three poınter altımeter As a consequence the Kollsman Instrument Dıvısıon of the Square D Company ıs now developıng such an altımeter Two other ıtems of avıatıon equıpment currently beıng developed by the Aır Force an absolute (radıo) altımeter and an aırborne dıstance measurıng devıce, are also ıncorporatıng thıs type of ındıcatıon

TABLE 31 2

Frequency of Various Types of Error Made by 97 USAF Pilots and 79 College Students in Reading the Conventional Three Pointer Altimeter

Description of Error		*Per Cent of Total Readings in Which Error Appeared*	
		Pilots	College Students
Reading to nearest numeral instead of to lower adjacent numeral— (Failure to consider more sensitive pointer)	100 Ft	0 09	0 11
	1 000 Ft	2 58	1 48
	10,000 Ft	1 72	2 11
	Total	4 39	3 69
Reading to lower adjacent numeral when nearest numeral is correct— (Failure to consider more sensitive pointer)	100 Ft	0 0	0 0
	1,000 Ft	0 26	2 22
	10 000 Ft	0 0	0 11
	Total	0 26	2 32
Displacement of digit in number series— (Interchange of digit with adjacent zero)	20 Ft	0 17	0 42
	100 Ft	0 86	0 95
	1,000 Ft	2 06	2 64
	10,000 Ft	0 86	1 48
	Total	3 95	5 48
Misreading of scale or numeral—	20 Ft	3 09	2 64
	100 Ft	1 20	1 05
	1,000 Ft	1 46	2 85
	10,000 Ft	0 09	0 53
	Total	5 84	7 07
Omission of one pointer—	100 Ft	0 0	0 0
	1 000 Ft	0 86	0 21
	10,000 Ft	0 86	1 05
	Total	1 72	1 27
Pointer exchange—	100 and 1,000 Ft	0 17	0 84
	100 and 10 000 Ft	0 0	0 0
	1,000 and 10 000 Ft	0 09	1 48
	Total	0 26	2 32
Repetition of reading on one pointer—		0 95	0 84
Complex and unclassified errors		0 86	1 48

altimeter) More significant perhaps, is the finding that only 0 7 per cent of the read ings erred by more than 1000 feet Most of the errors in reading indicator design D resulted from assigning 10 feet instead of 20 feet to each of the graduation intervals between numerals

Indicator design E The substitution for 2 of the pointers on the altimeter of 2 rotating dials appearing through a window appears to have no advantage This indi cator was designed so that under most cir cumstances only 1 number would appear on each of the 2 rotating dials But if such dials rotate continuously (rather than in termittently) during altitude changes, as assumed in this test, it is inevitable that at certain settings 2 numbers will be equally visible Such indications are very difficult to read correctly

Indicator design F On this indicator the range covered by the indicating pointer and scale is dependent upon range limits shown in the window The high proportion of errors and slow reading time suggest that the required changes in interpretation of the primary scale are difficult for human beings to carry out

Indicator designs G and H The vertical moving scale instruments proved to be easy to read in this experiment The virtues of such instruments for check reading and qualitative reading were not evaluated in this study

Indicator design I This indicator, which simulates a simple Veeder counter, was read with greatest speed and accuracy of all the indicators This would suggest that where only quantitative readings are to be provided this would be the most desirable type of instrument It is believed that for check reading and for qualitative reading such an instrument would be inferior to one using a moving pointer

Types of Error in Reading the Three Pointer Altimeter

Because of the widespread use of the 3 pointer altimeter, and because of the fre quent use of this type of multiple pointer indication for other purposes it seemed worth while to make a more detailed an alysis of the types of errors made in quan titative readings of this instrument This analysis was based on the same data that have already been summarized in Table 31 1 and Figure 31 1 It will be recalled that 97 USAF pilots and 79 college stu dents each made 12 readings on the 3 pointer altimeter This gave a total of 1164 readings by pilots and 948 by non pilots

The detailed classification of errors into types and sub types is shown in Table 31 2 along with the per cent of total readings in which each occurred Two or more types or sub types of errors were in some cases charged against a single erroneous read-ing For this reason the figures in the per cent columns total up to more than the total per cent errors as reported in Table 31 1 For detailed descriptions of all the error types, and the assumed mental proc esses which led to the incorrect answers, the reader is referred to Aero Medical Lab oratory Memorandum Report No MCREXD 694 14A

Discussion

In an experiment such as this a number of questions arise with regard to the suit ability of the criterion measures which have been used and with regard to the effect of the subject group upon the results For this reason there have been included in Table 31 1 a number of correlation coef ficients which bear on these problems

A serious question facing the experi menter is the effect of the experience of the subject group upon the validity of the findings In the present experiment two subject groups were used which rep resented extremes in experience as related to the task being performed All USAF pilots had had considerable experience in reading one of the experimental indicator designs along with general experience in reading aircraft instruments The college students, on the other hand, included no pilots or other military air crew members In spite of this difference in background of experience the two groups gave highly similar results as indicated by a correlation between the two groups of 95 on per cent errors and 99 on seconds per reading This would suggest that the previous experience

of the subjects is of relatively minor importance in an experiment of this type

In the present experiment neither speed nor accuracy of response were controlled, thus making possible two independent criteria for evaluation of the different dial designs In Table 31 1 the correlations between speed and accuracy for the different dial designs are 91 for pilots and 95 for college students, indicating very high agreement between the two criteria for goodness of the several indicator designs Or stated differently, the indicator designs which were read most rapidly were also read most accurately Correlation coefficients between speed and accuracy of individuals for all designs are also positive, but much lower, 38 for pilots and 44 for college students These values indicate, however, that in general the individuals who read the indicators most rapidly also read them most accurately In a previous study by the author (2) on clock dial designs the correlation coefficients were likewise positive, but somewhat lower in magnitude

In two previous experiments on instrument design by Loucks (4) and Sleight (5) a somewhat different technique was used in that the instrument exposure time was controlled tachistoscopically and only accuracy of reading was measured Such a technique might be expected to force an increased error rate and thus accentuate the differences between indicator designs It is the belief of the author, however, that such a control of exposure does not constitute control over response time, but serves rather to restrict the number of visual fixations of the displayed material The actual response may be delayed for several seconds during which the subject retains a mental image of the indicator scale and pointer

It is quite possible that in the experiment of Sleight (5) the use of a controlled exposure time which did not permit a change in the preparatory eye fixation led to erroneous findings It is believed that this technique favored the fixed pointer indicators on which the subject was able to anticipate the location of the pointer The 2 fixed pointer indicators in the present study, designs G and H, showed no general superiority over the only comparable moving pointer indicator, design D

Summary

An evaluation was made of the speed and accuracy with which quantitative readings could be made of 9 experimental altitude indicators The results are considered to apply also to other types of quantitative indication which require very great scale length Evaluation of the various indicator designs was made by having 97 USAF pilots and 79 college men read 12 settings on each instrument The instrument faces were reproduced in test booklets which provided spaces for writing in the readings Both accuracy and speed of reading data were obtained for each of the nine indicator designs

The major conclusions indicated by the results of this investigation are as follows

1 The combining into a single numerical value of the indication from 2 or more pointers, or from a pointer and rotating subdials, is a relatively difficult task for human beings Such instruments are conducive to very large errors in reading

2 The ease with which long scale indicators can be read quantitatively appears to depend upon the extent to which the digits are already combined in the proper sequence by the instrument

3 A multiple pointer instrument such as the altimeter with continuous motion of the non sensitive pointers is frequently read too high by a complete revolution of the sensitive pointer

4 The speed and accuracy of instrument reading are positively correlated, indicating that gains in reading speed can normally be expected to improve accuracy also

5 College men without altimeter reading experience showed virtually the same pattern of results in this study as highly experienced USAF pilots, suggesting that instrument reading difficulties are quite basic in nature and not readily modified by experience

References

1 Fitts, P M and Jones, R E, "Psychological Aspects of Instrument Dis

play I Analysis of 270 pilot error Experiences in Reading and Interpreting Aircraft Instruments USAF Air Materiel Command Memorandum Report No TSEAA 694 12A, 1947

2 Grether, Walter F, "Factors in the Design of Clock Dials which Affect Speed and Accuracy of Readings in the 2400 Hour Time System, *Journal of Applied Psychology* 1948, Vol 32, 159–169

3 Grether, Walter F, Designing Instrument Dials for Quick, Accurate Reading, *Machine Design* 1948, Vol 20, 150–152 and 208–209

4 Loucks, R B, Legibility of Aircraft Instrument Dials The Relative Legibility of Tachometer Dials, AAF School of Aviation Medicine Project No 265, Report No 1, 1944

5 Sleight Robert B, The Effect of Instrument Dial Shape on Legibility," *Journal of Applied Psychology,* 1948, Vol 32, 170–188

Types of Errors in Location Judgments on Scaled Surfaces II Random and Systematic Errors *

ADELBERT FORD

This research was executed under Contract No W28 099 ac 130 between the Institute of Research Lehigh University, and the USAF Air Materiel Command, Watson Laboratories Red Bank, N J The investigation was made to ascertain the accuracy of radar operators in the interpretation of scope signals

A large variety of instruments require operators to report the position of a signal, such as a white spot, by reading its position with reference to superimposed scaling lines In dealing with types of radar associated with the navigation of aircraft a single large error could cause loss of life and the destruction of expensive equipment

In the present article, using the same scaling and problem sequences, we propose to show (1) the size of the *random errors* caused by the limiting effects of interpolating scale values of specific scales, and (2) certain systematic errors consisting in particular of the *confusion error,* defined as a mistaken interpretation of the numerical value of the scale points, and what we shall call *persistence errors* defined as a proneness of some subjects to bias reports in a sequential series by memory effects of the previous reports

Although the present report is specifically concerned with position reporting from scaled areas, it will probably be instantly perceived that some of the principles are perhaps equally applicable to linear scales The consequences of this error analysis are much more basic than the narrow application to radar scopes

Fineness of Scaling and Random Error [1]

As illustrated in a previous article, there were three types of scaling used for these experiments (1) a scope with a zero line of reference across the field but no other scaling assistance other than a sample scale printed on the side of the scope for comparison, (2) a scope with a so called 100 foot Reference Line' located parallel to and 0 4 inch away from the zero line of reference, and (3) a scope with a multiple system of parallel lines, separated by tenths of an inch, each line representing 25 scaled feet

* Reprinted from *Journal of Applied Psychology,* Vol 33, No 4, August 1949

[1] Readers who possess a cleared status for restricted reports will find a more elaborated description of the tables and calculations in A Ford and M H Getz Types of Errors in the Reading of GCA Scaled Scopes Technical Report No 4, Contract W28 099 ac 130, Watson Laboratories, Air Materiel Command, USAF, 31 August 1948 Restricted

For practical reasons the errors were all reduced to percentage values in this section of the data, using only pips which were 50 or more scaled feet from the zero line of reference Figures 32 1 to 32 3 are based on the composite records of five subjects (It will be shown later that individual differ ences in random error are small)

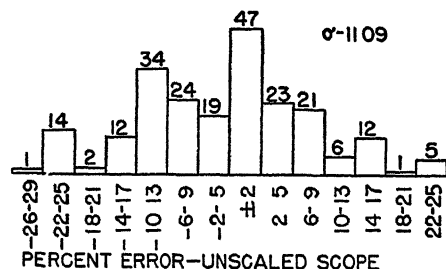



FIGURE 32 1 *Distribution of errors on the unscaled scope*

Figure 32 1 shows that for the unscaled scope, the standard error was 11 09 per cent of the space being estimated Figure 32 2 shows that the use of side lines, 0 4 inch away from the zero line of reference, reduced the standard error to 8 48 per cent Figure 32 3 shows that with the use of a multiple system of lines, one tenth inch apart, the standard error is now reduced to 4 59 per cent

Now Garner (1) has shown that on

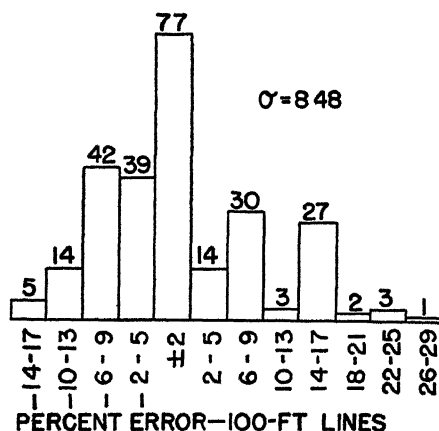



FIGURE 32 2 *Distribution of errors on the scope with 100 ft reference lines*

PPI type scopes, with scaling in the form of concentric rings, scaling of the degree of fineness in our multiple system pro duced confusion errors, decreased accuracy and promoted longer reaction times We shall substantiate Garner s statement with respect to confusion errors, but we shall have to indicate, from evidence in Figures 32 1 to 32 3, that the smallest spread of

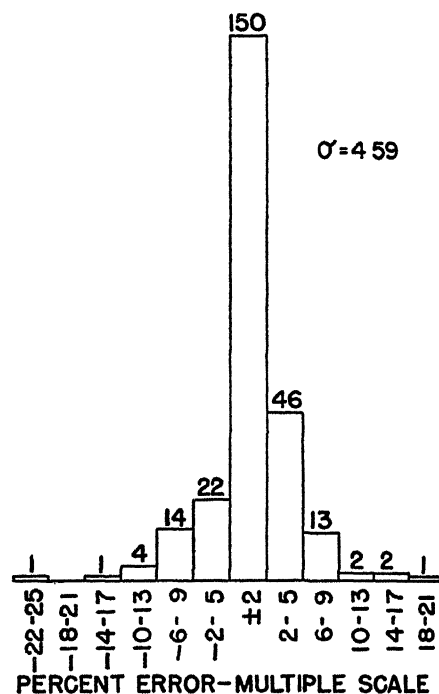



FIGURE 32 3 *Distribution of errors on the scope with multiple scaling*

random error was produced for the finer scaling We found no statistically reliable difference in verbal reporting reaction time This may be a difference between human reactions on polar scaling, which Garner used, and rectangular scaling, which we used

At this stage in the experiments we went into a more detailed gathering of data on the finely scaled scopes, to see whether or not the advantage of a smaller random error was not offset by the presence of systematic errors which could not be tol erated

Absolute Amount of Random Error

Since we have ascertained that the more finely scaled scope yielded the smallest random error in percentage figures, we shall now confine our measurements to the absolute values in this scaling situation (lines in tenths of an inch, representing 25 scaled feet of elevation with 100 foot lines emphasized)

In Tables 32 1 and 32 2 the standard deviation of the error spread is computed omitting the confusion errors around the 100 foot scaling line, which are obviously not random Mistaken numerical interpretations around the 25 foot scaling line can not be distinguished easily from random errors but we shall make an attempt, later, to show they exist by statistical analysis

Individual differences, for untrained subjects on group experiments, with clear, uniform signals, are presented in Table 32 1 It appears safe to say, from these data, that average intelligent operators should be able to report elevation deflections to a standard error of a plus or minus 0 020 inch of scope distance, under such conditions This represents an error in judging the elevation of a plane of 5 or 6 feet, presumably trivial Trained subjects are much more nearly alike in error spread and we have combined the runs in Table 32 2 to show the absolute error under 6 different experimental conditions

The 6 conditions in Table 32 2 are as follows

Condition A Five trained subjects Individual experiments Artificial scope with clear uniform signals Rectangular presentation Single task elevation reporting

Condition B Nineteen untrained subjects Group experiments before a large screen Same problem materials as Condition A Rectangular display Single task elevation reporting

Condition C Five trained subjects Individual experiments Artificial scope with clear uniform signals Sector presentation Single task reporting

Condition D Nineteen untrained subjects Group experiments before a large screen Same problem materials as in Condition C Sector display Single task reporting

Condition E Six trained subjects Individual experiments Simulator reproductions of field radar Typical pip variations in contour, size, brightness, shape, and hazy

TABLE 32 1

Random Error Standard Deviation, Group Experiments, Individual Differences

Subject	*Stand Dev in Scaled Feet*	*Stand Dev in Scope Inches*	*Number of Readings*	*Subject*	*Stand Dev in Scaled Feet*	*Stand Dev in Scope Inches*	*Number of Readings*
M J D	2 8	011	89	R K S	4 9	019	89
N J R	4 1	016	89	C A W	5 2	021	87
J H J	4 5	018	114	C E F	5 5	022	90
I E K	4 5	018	90	M K S	5 6	022	88
R B T	4 5	018	89	P A W	5 8	023	86
A H F	4 6	018	88	K M	6 0	024	110
W J K	4 6	018	64	M S W	6 1	024	88
A W R	4 7	019	89	D L H	6 5	026	88
A P R	4 8	019	63	B J J	7 0	028	87
M B C	4 9	019	90				

Note Confusion errors at the 25 foot minor scaling line cannot be accurately separated from random errors The above standard errors include these, and are probably all too large See Table 32 3 for an attempt at separation

In this table the subjects are arranged in the order of best to worst and all are untrained The scaling consists of the multiple system with lines a tenth of an inch apart

edges Sector display Single task reporting

Condition F Six trained subjects Individual experiments Simulator reproductions of field radar Same problem materials as Condition E Sector display Double task reporting, alternating elevation reports with range reports

The standard deviation of error distributions appears at the base of each column in Table 32 2, expressed both in scaled feet and in inches of actual scope distance

Conditions A, B, C, and D all involve artificial scope pictures with clear, uniform signals The conclusion that an average operator should be able to interpret distances, under these conditions, to a standard error of a plus or minus 0 020 inch is again substantiated If a radar scope could be designed with such clear and uniform pips, and using scaling of this degree of fineness, this gives the human expectancy

Condition E, using reproductions of an actual radar scope, shows that the random error is about doubled, due to signals which vary in shape, size, intensity, haziness of edges, etc In the artificial series the reports were 10 seconds apart In this simulator series the operator reported every tenth pip, with the scan line crossing the

TABLE 32 2

Distributions of Errors under Various Conditions, Elevation Reporting, Multiple Scaling, All Subjects Combined

Error Scaled Feet	*Character of Run*						*Location of Types of Errors*
	(A)	(B)	(C)	(D)	(E)	(F)	
+110		1					Approximate band of confusion errors
+105		1		4			around the 100 foot major scaling
+100	3	8	6	10			line Errors of overestimation
+95		6		3			
+90				1			
+85				1			Approximate band of confusion errors
+80		1					around the 75 foot minor scaling
+75				1			line Errors of overestimation
+70							
+65							
+60							Approximate band of confusion errors
+55						1	around the 50 foot minor scaling
+50		1					line Errors of overestimation
+45						1	
+40		1					
+35				1	1	10	Approximate band of confusion errors
+30					7	19	around the 25 foot minor scaling
+25	3	1		6	13	29	line Errors of overstimation
+20		4	3	9	38	97	
+15	2	7	2	7	106	135	Central band of random errors
+10	56	67	76	58	159	174	
+5	268	370	375	365	262	193	
00	906	902	688	774	414	189	
−5	236	240	366	366	275	211	
−10	32	34	83	75	170	153	
−15	13	8	9	10	64	111	

For a description of the character of each run, as designated by A, B, C, D, E, and F, see pp 209–210 of the text

TABLE 32 2 (Continued)

Error Scaled Feet	*Character of Run*						*Location of Types of Errors*
	(A)	(B)	(C)	(D)	(E)	(F)	
−20	2	7	4	3	32	77	Approximate band of confusion errors around the 25 foot minor scaling line Errors of underestimation
−25	3	7		3	19	27	
−30		4	1		3	13	
−35			1		2	6	
−40		1				4	Approximate band of confusion errors around the 50 foot minor scaling line Errors of underestimation
−45							
−50						1	
−55		1			1		
−60							
−65							Approximate band of confusion errors around the 75 foot minor scaling line Errors of underestimation
−70							
−75				1			
−80							
−85							
−90	1						Approximate band of confusion errors around the 100 foot major scaling line Errors of underestimation
−95		2		3			
−100	2	2		3			
−105				1			
S D Feet	4 3	5 0	5 1	5 3	9 8	13 3	
S D Inches	02	02	02	02	04	05	

scope once every second Rate of reporting was approximately the same, therefore

Condition F is just like Condition E, except that the operator had to keep his attention on two tasks in alternation, elevation reporting and range reporting The increase in standard error, from 9 8 feet to 13 3 feet, represents the effect of giving an operator an additional task It may be presumed that the more tasks the radar operator is required to do simultaneously the less accurate he will be on each This conclusion may seem to be something like proving the obvious, but it must be remembered that there is a proposal to make one man do what was previously done by from 3 to 5 men on GCA radar installations The need for 1 man operation is urgent, and the present study is merely an attempt to show that multiple tasks must be accompanied by extreme work simplification, if we are to avoid intolerable reporting errors One confusion error, of the amount shown in Table 32 2 at the 100 foot line, could wreck an air transport

Figure 32 6 shows the fit of a normal curve of distribution to the actual error distribution for the data of Condition E, reproductions of actual radar scopes

Confusion Errors

Scales, both linear and surface types, consisting of major lines with numerical values, and minor divisions which are supposed to assist in interpolation, are subject to mistaken interpretation of figures and errors in counting division points

Table 32 2 shows a clear existence of mistaken interpretation at the 100 foot value This is verified by subjective reports, many times The 100 foot line is called a 200 foot line, or the line of zero reference is mistaken for a 100 foot side line There was no case of an error as great as 200 feet, but it was theoretically possible

Also, at the 75-foot, the 50-foot, and the 25-foot distances there is an equal probability of assigning wrong numerical interpretations. These are fairly clear at 50 feet and up. Unfortunately the confusion errors at the value of 25 feet overlap with the

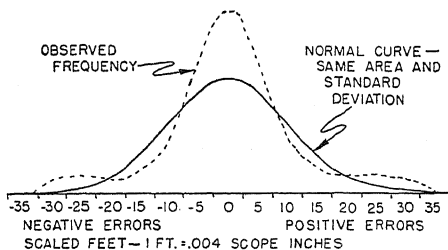



FIGURE 32.4. *Type of fit for normal curve when errors of 25 ft., the position of a minor scale division, have been included.*

curve of random error. In fact, there is no way of separating confusion from random errors, at this position, but there may be a statistical way of showing facts which support the belief that they must be there.

Assuming that random error distributions should approach the curve of normal probability, an hypothesis which has consider-

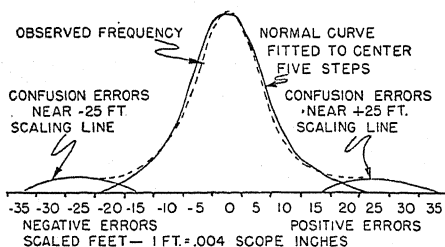



FIGURE 32.5. *Hypothetical improvement of normal curve fit when errors at the 25 foot scaling position have been excluded. Presented to explain the* X^2 *improvement shown in Table 32.3.*

able support, and that systematic errors will cause typical and expected distortions from normalcy, we may resort to the X^2 test for these data. And in this use of the Fisher technique, it isn't just the bald fact that a misfit has occurred, but *where in the curve* the misfit is found, whether or not it is over the values which correspond to the minor or major scale points, that should prove of interest in spying out the presence of confusion errors mixed with random errors at the 25-foot distance.

Figure 32.4 shows the typical result we get when we try to fit a normal curve on our error distributions. The normal curve is plotted using the standard deviation of the distance from −35 to +35 feet, which includes confusion errors around 25 feet.

The X^2 test always resulted in *too many errors* over the 25-foot position, and the discrepancy was *always positive* for every

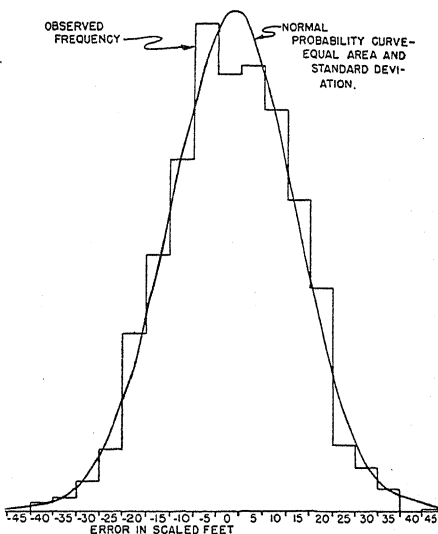



FIGURE 32.6. *Normal curve fit—random errors. Elevation reports in double task experiments.*

distribution beginning with Condition A through and including Condition E. This always produced the appearance of a leptokurtic hump at the center.

Figure 32.5 shows our hypothesis of what would happen if we determined the standard deviation by the central band of random error, only, and deliberately assumed that the excess of readings over the 25-foot point is due to confusion errors, not random errors.

Therefore, we adjusted the standard deviation value to fit the central band of error values, from −15 to +15 feet, and

TABLE 32 3

Artificial Scope Runs

x^2 Tests of Curve Fit for a Normal Distribution of Error, Central Band of Random Error

Condition	*Curve Area of Central Band*	*Stand Dev −35 to +35 feet*	*x^2 Fit Central Band*	*Stand Dev − 15 to +15 feet*	*x^2 Fit Central Band*	*Number of Readings Central Band*
(A)	98 5%	4 31	167 25	3 40	80 19	1499
(B)	98 8%	5 10	24 28	4 65	5 89	1588
(C)	97 7%	5 01	225 97	3 70	48 67	1613
(D)	97 7%	5 30	97 01	4 40	9 92	1638
(E)	81 8%	9 80	39 79	8 50	23 27	1280
(F)	——	13 33	1 78	——	——	——

NOTE The central band for Condition F was from −35 to +35, and was divided into five step intervals No attempt was made to improve the fit because this was already as good as could be obtained The area of this central band was 99 5% of the total distribution

In this table the errors from −15 feet to +15 feet are hypothetically considered as being the band for pure random error (see Table 32 2) and that this region should fit a normal probability curve The fit to the central band is tried two ways with the supposed confusion errors included, and with the confusion errors excluded i e by computing the standard deviation only on the central band

applied the X test again The differences between the two assumptions are shown in final X answers in Table 32 3 Without exception, for the artificial scope series A to D inclusive, a X fit for 98 per cent of the readings was greatly improved The astonishing thing was the discovery that the distribution for Condition F, double task reporting, was already almost a perfect normal curve, and could not be improved

TABLE 32 4

Distributions of Errors

Range Reports

Reproductions of GCA Field Radar Scope

Error in Scaled Miles	*Initials of Subjects*							
	R C	D M	D M S	J H F	W A S	C B	*Total*	
+1 0					1		1	Band of
+ 9					2		2	persistence
+ 8								errors
+ 7								
+ 6					1		1	
+ 5				1	1		2	
+ 4								
+ 3								Band of ran
+ 2								dom and con
+ 1	13	29	17	31	32	17	129	fusion errors
0 0	99	127	118	127	123	106	700	
− 1	84	50	68	46	49	70	367	
− 2	14		5	2		13	34	
− 3						1	1	

by any assumptions of systematic distortion

We are inclined to believe, therefore, that the approximate bands for the regions of confusion errors in Table 32 2 are essentially correct This means that, in reducing the random error by more finely divided scales, we have introduced an intolerable numerical confusion error, extremely dangerous for the practical navigation of aircraft by ground control radar Therefore, no recommendation is made to use such a scale More simplified methods of signal tracking must be designed, especially for one man operation

Persistence Errors

A rather broad definition of a persistence error may be It is the tendency of an operator to bias a present report because of the mental persistence of a previous report

We uncovered the existence of this possibility through two subjects whose data are plotted in Table 32 4 The first evidence was a sort of verbal stereotyping occurring when operators had to attend to two things alternately Table 32 4 is drawn from the double task experiments of Condition Γ

An operator would be reporting consecutive range values, ' 6 point 2, 6 point 1, 6 point zero, and when he passed into the 5 mile zone he went on, 6 point 9, 6 point 8," and then suddenly remarked, Oh, I meant 5 point 8 This is essentially the situation for Table 32 4

This led us to wonder if something similar to this might not have been happening, to susceptible subjects, in the previous elevation serial reporting Therefore, we computed the algebraic mean of errors following larger previous values, and subtracted this from the algebraic means of reports following smaller previous reports This difference is susceptible to a calculation for *reliability of differences of means* Table 32 5 shows the results of this survey Although only 6 out of 26 subjects showed a significance of difference better

TABLE 32 5

Trend of Algebraic Mean Error in Relation to Previous Report
Elevation Scale

Subject	Difference	Subject	Difference	Subject	Difference
		1 *Group Experiments Artificial Scope*			
M B C	+*2 38*	W J K	+ 32	N J R	+ 79
A H F	− 24	L E K	+ 18	R K S	+*2 05*
C E F	+ 45	K M	+1 56	R B T	+*1 91*
D L H	+ 88	A W R	− 61	C A W	−1 37
B J J	+1 18	A P R	− 28	M S W	+ 21
J H J	+1 07				
		2 *Individual Experiments Artificial Scope*			
L A A	+*1 05*	F P H	− 31	D M S	+ 84
W A S	+1 02	R J R	+*1 01*		
		3 *Simulator Reproductions of Field Radar*			
C R B	+*5 30*	J H F	+3 07	W A S	+4 60
R C	−2 24	D M S	+ 80		

A plus sign means that the subject tended to veer his reports in the direction of the preceding report A minus sign means that the subject tended to bias away from the preceding report The calculation is the difference in means where the preceding report was higher as compared with readings where the previous report was lower Figures in italics are better than the 1 per cent level of significance Differences are in scaled feet

than the 1 per cent level, the general preponderance of plus values (20 out of 26) may carry some weight

Granting that some subjects are susceptible to this effect, the size of the error trend is actually too small to be of any serious consequence for the practical control of aircraft A biasing effect of 2 feet, or even 5 feet, would not be intolerable On range reporting it is conceivable that a mistake of 1 mile might be serious

SUMMARY

1 The use of finer scaling, with minor scale division to tenths of an inch viewed at 16 inches, reduces random errors to a standard deviation of a plus or minus 0 020 inch of scope distances, for clear uniform pips, and 0 040 inch of scope distance for reproductions of actual radar pips

2 The introduction of this finer scaling produces a proneness for confusion errors, defined as misinterpretation of the numerical values of scale positions These errors may reach such a size as to endanger the navigation of an aircraft being guided by such operating reports

3 Requiring an operator to alternate between two tasks in rapid succession has the effect of increasing the size of the random error, in our situation, about 30 per cent

4 Some subjects have a tendency to bias each report in a series by the mental persistence of the previous report Only a minority of subjects do this consistently, and the amount is relatively small for practical significance

5 Fine scaling, for one or more variables, is not recommended on the basis of present data for radar scopes

REFERENCES

1 Garner, W R Some Perceptual Problems in the Use of VG Remote PPI, Report of Research under Contract with the Office of Research and Inventions U S Navy, 166–I–2, 15 September 1946 Restricted The Johns Hopkins Psychological Laboratory p 34

2 Ford, A and Getz, M H, Types of Errors in the Reading of GCA Scaled Scopes, Technical Report No 4 Contract W28 099 ac 130 Watson Laboratories, Air Materiel Command, USAF, 31 August 1948 Restricted

Psychological Factors in Instrument Reading II The Accuracy of Pointer Position Interpolation as a Function of the Distance Between Scale Marks and Illumination *

WALTER F GRETHER and A C WILLIAMS, JR

This experiment was carried out at the University of Illinois by Dr A C Williams, Jr, under a dollar a year contract with the USAF Air Materiel Command Dr W F Grether proposed the study, designed and procured the necessary dials and prepared the present report The basic data have been presented previously in Army Air Forces Aviation Psychology Program Research Report No 19 Chapter 7, and in USAF Air Materiel Command Memorandum Report No TSEAA 694 1

The reader of instruments is normally expected to obtain values of greater precision than the graduations placed upon the instrument scale To accomplish this he must interpolate, that is, estimate the relative distance of the pointer from the two scale marks between which it falls and assign an appropriate value to this position The accuracy with which this can be done

* Reprinted from *Journal of Applied Psychology* Vol 33, No 6, December 1949

obviously limits the precision with which any given scale can be read The accuracy of such interpolation, moreover, will be influenced by several variables in the scale design and the conditions of reading For a prediction of reading precision obtainable with different instrument designs under various conditions of viewing the effect of the significant variables must be known In the present experiment the accuracy of pointer position interpolation was studied as a function of (a) diameter of the dial, (b) angular separation of the scale divisions, and (c) simulated day versus night viewing conditions It will be shown in the presentation of the results that the first two variables can be reduced to a single one, namely, the length of the arc (in visual angle or inches) between scale marks

A problem in scale design to which the present investigation is particularly relevant is concerned with the question of how finely a scale should be divided in order to provide maximum reading accuracy In an investigation by Loucks (5) the legibility of tachometer dials was investigated using rather short exposure (0 75 sec) For three dials with graduations of 100, 50, and 20 RPM respectively the percentage of reading errors increased as the value and size of the graduations decreased From this finding it might be concluded that placing the graduations rather close together will decrease rather than increase reading accuracy The findings of Kappauf, Smith and Bray (3), and Kappauf and Smith (4), in experiments where the exposure interval was not limited, disagree with those of Loucks In their experiments dials graduated in units gave greater reading accuracy and speed than dials of the same size but graduated in 5 or 10-unit steps However, the superiority of 1-unit over 5 unit graduations was rather small and not at all in proportion to the increased number of graduation marks These latter results are in agreement with those of an investigation by Grether (1) on the reading of clock dials With 1 minute as the criterion of reading accuracy, dials with 1 minute graduations gave higher reading accuracy than similar dials with only 5 minute scale marks

A possible explanation can be offered for the discrepancy between the findings of Loucks (5) and later investigators It is quite probable that as the number of scale marks is increased more eye fixations are required to make each reading By limiting the exposure time and consequently the number of eye fixations Loucks may have favored those dials with more widely spaced graduations

In the study of scale designs it is helpful to distinguish between two general types of errors encountered in dial reading studies There are first the precision errors or errors of interpolation These can never exceed in magnitude the value of the smallest interval on the scale The other type may be called comprehension of interpretation errors, in which an incorrect value is assigned to the graduation mark against which the pointer is being read Comprehension errors are frequently very large and are usually some multiple of the minor, intermediate, or major scale divisions In a study by Grether (2) of altimeter reading, for example, most of the errors were of this latter sort, with errors of 1000 feet being particularly common It is important to recognize that many of the dial reading studies up to the present have been concerned only with the interpolation type of errors, when in actuality the larger comprehension errors are far more serious in practical instrument reading situations It is quite possible that the presence of a large number of graduation marks on a dial may greatly increase the probability of large comprehension errors and thereby nullify the precision which a finely graduated scale makes possible

The spacing of the divisions on a scale is usually not a variable concerning which the instrument designer has a free choice Normally the physical length of the scale, the range of values to be covered, and the desired accuracy of reading are fixed by the particular application Based upon these requirements the designer must then select values (usually 1, 2, 5, or decimal multiples of these) for his scale increments which will give him reasonable spacing between graduation marks

The aim of the present study was to provide the instrument designer with data

from which to predict how the accuracy of readings will be affected by the physical length of the interval into which he sub divides a scale Measurements were made under two lighting conditions comparable to those under which aircraft instruments are viewed Emphasis in this study was placed on interpolation errors The more complex comprehension type of errors were recorded but were relatively few in number and were not subjected to analysis

Apparatus

For the purpose of this experiment a series of 16 simulated instrument dials was prepared A sample dial and pointer are shown in Figure 33 1 Four sizes of dials were used as follows 1, 1⅞, 2¾, and 4 inches in diameter The particular dimen

FIGURE 33 1 *Sample dial and pointer (1⅞ inch diameter and 20 degree angular sep aration between scale marks*

sions of the two intermediate sizes were chosen to duplicate standard aircraft instru ments Each size of dial was produced with four different graduation intervals, defined in terms of the angular separation between scale marks, as follows 5, 10, 20, and 40 degrees Except for the variations in di ameter and size of graduation intervals, all dials were identical The intermediate graduation marks were ⅛ inch in length and approximately 0 02 inches in width The major graduation marks at each end of the scale were the same width but ¼ inch in length The numerals on all dials were ⅛ inch in height All pointers were 3/32 inch in width and of such a length that the tip reached to the inner edge of the shortest graduation marks All dials covered a range of from 0 to 50 units as shown in Figure 1 with graduation marks only at the 0, 10, 20, 30 40, and 50 posi tions and numerals only at 0 and 50 These dials were engraved on brass plates, which were then painted a flat black and the engraved markings filled with yellow fluorescing paint (pale yellow in daylight) as used on the latest type of USAF instru ments

The experimental dials were presented singly in a panel opening 30 inches from and perpendicular to the subject s eyes Daylight conditions were simulated with a fluorescent type daylight lamp which provided an illumination of 45 foot candles at the panel opening For simulation of night conditions the subject s room was completely darkened and the dial illumi nated with a standard C 5 ultra violet air craft instrument panel light operating at maximum intensity No means were avail able for obtaining a quantitative measure ment of the brightness of the scale markings under ultra violet illumination Covering the opening in which the ex perimental dials were presented was a mechanical shutter operated by the experi menter

On the experimenter s side of the test panel was a carriage on which four of the dials could be mounted side by side This carriage rode upon two horizontal tracks parallel to the screen To present any one of the dials the experimenter moved the carriage so that the desired dial would appear in the panel opening At the ex perimenter s side of the carriage were 4 master setting dials 5 inches in diameter On each of these dials was a pointer con nected to the same shaft as the pointer on the dial to be read by the subject On the experimenter's dials were closely spaced graduations which made possible accurate

settings to one tenth of the space between graduations on the subject s dials

Also provided at the experimenter's station was a lever for manual operation of the shutter used to expose the dial to the subject This lever was used also to operate an electric timer through a suitable switch Thus, the timer indicated the time during which the shutter remained open Since the experimenter closed the shutter as soon as the dial reading had been completed, the reading on the clock gave a crude measure of the reaction time on each test trial Several other methods of measuring reaction time were tried but found to be unsatisfactory

Eighty male college students were used as subjects in this experiment Only men with 20–20 binocular vision (corrected or uncorrected) were accepted The subjects were seated in a chair in front of the screen with their eyes 30 inches from the panel opening and with the line of sight perpendicular to the panel opening in order to eliminate parallax The subjects were divided into groups of 20, each group being tested on a set of four dials The four dials included one of each diameter and one of each graduation interval Each subject was given a total of 80 trials, equally divided among the four dials in a random sequence Of each group of 20 subjects, 10 were tested under simulated daylight conditions and the remaining 10 under simulated night conditions

A variety of dial settings were chosen so as to represent all portions of the dial from 0 to 50 The actual numbers to be read were the same for all dials although the order of presentation was randomized The subjects were instructed to read the dials as quickly and accurately as possible to the nearest whole number As can be seen in Figure 33 1, the reading to the nearest whole number required estimation to the nearest one tenth of the distance between graduations

On each trial the experimenter set the pointer of the dial to be presented, then opened the shutter and waited for the subject s verbal response, following which the shutter was closed and the subject s reading and the clock score recorded

Results

The experimental design resulted in 200 readings on each of the 16 specific dials under each of the lighting conditions For each reading both error and time data were obtained The error data consisted of the deviations of the readings from the actual settings These deviations could be either negative or positive and increased in step intervals of one, or one tenth of the space between graduation marks For purposes of analysis, however, error distributions were made without regard to sign Since the distributions of these errors, and also response times, were considerably skewed, with the modal error being zero for many of the dials, the statistical treatment presented in this report is limited to medians and 75th percentiles Means were computed for all the data and found to present the same general picture as the medians, but the values were inflated because of the skewness

A summary of the error data for the 16 dials is shown in Table 33 1 In the third column, it will be noted, the two variables of dial diameter and angular spacing of the divisions have been reduced to a single variable, namely, the length of graduation interval defined as the arc between the inner ends of the shortest scale marks This value can be described also as the distance the pointer tip must travel between adjacent graduations

This length of the inner arc is presented in inches with the multiplying factor provided for conversion to minutes of visual angle for the 30 inch viewing distance A comparison of the columns of median errors for daylight and night conditions in Table 33 1 reveals no consistent difference between these two lighting conditions Only for the dial with the most closely spaced divisions does the performance under daylight appear to be superior For this reason it seemed safe to combine the two sets of error data in the remaining columns of the table

Some of the most important findings contained in Table 33 1 are presented graphically in Figures 33 2 and 33 3 With length of graduation interval along the base line, the median and 75th percentile errors

TABLE 33 1

Summary of Data on Accuracy of Pointer Interpolation as Function of Dial Diameter and Spacing of Scale Divisions

Dial Diameter inches	*Angular Spacing degrees*	*Length of Inner Arc inches**	*Median Error Daylight % of interval*	*Median Error Night % of interval*	*Median Error Combined % of interval*	*75th Percentile Error Combined % of interval*	*Median Error Combined degrees*
1	5	032	21 8	31 0	26 4	45 5	1 32
1	10	065	17 8	18 8	18 3	28 5	1 83
1	20	130	14 3	14 9	14 6	21 2	2 92
1	40	261	13 2	12 1	12 6	17 6	5 04
1⅞	5	070	20 0	19 4	19 7	31 8	0 99
1⅞	10	141	11 2	14 2	12 8	18 5	1 28
1⅞	20	238	12 4	10 3	11 4	17 1	2 28
1⅞	40	567	7 8	8 1	8 0	13 9	3 20
2¾	5	109	15 4	15 7	15.5	21 2	0 78
2¾	10	218	12 0	9 3	10 6	16 5	1 06
2¾	20	436	9 4	9 1	9 3	15 2	1 86
2¾	40	872	8 1	8 1	8 1	14 1	3 24
4	5	163	14 9	13 9	14 4	20 0	0 72
4	10	327	9 7	9 0	9 3	15 2	0 93
4	20	654	8 8	7 9	8 3	14 3	1 66
4	40	1 309	9 8	9 1	9 4	15 0	3 76

* For conversion to minutes of visual angle multiply by 114 7

TABLE 33 2

Median Time for Interpolation of Pointer Position as a Function of Dial Diameter and Spacing of Scale Divisions

Dial Diameter inches	*Graduation Interval*			
	5	10°	20°	40°
	Seconds per reading for daylight conditions			
1	1 98	1 78	1 91	1 84
1⅞	1 73	1 80	1 86	1 76
2¾	1 83	1 85	1 73	1 77
4	1 87	1 75	1 68	1 90
	Seconds per reading for night conditions			
1	2 48	2 26	2 10	2 13
1⅞	2 18	2 14	2 02	2 10
2¾	2 15	2 05	2 02	2 06
4	2 00	2 02	2 08	2 23

are plotted as per cent of the interval in Figure 33 2 and as absolute values in Figure 33 3 Figure 33 2 will be recognized as a typical Weber function in which threshold ratios (DI/I) are plotted as a function of the stimulus intensity (I) In this figure, it will be noted the relative accuracy of interpolation is very nearly constant for graduation intervals above 0 5 inch, with a slight rise for the largest interval

with which the various points fit the curves in these two figures indicates that the combination of dial diameter and angular separation into the single variable of length of graduation interval was justified

The results of the measurements of response time in this experiment are summarized in Table 33 2 It is apparent that there are no consistent relationships between response times and the dial dimen

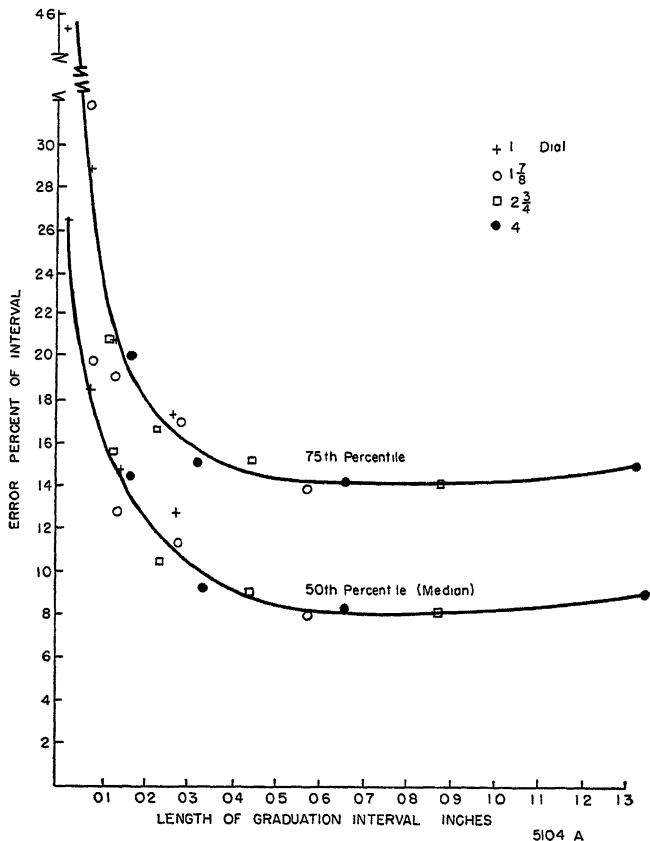



FIGURE 33 2 *Relative error of interpolation as a function of length of graduation interval*

In Figure 33 3 the absolute values for interpolation thresholds fall very nearly on a straight line, which if extrapolated to zero intervals would intercept the ordinate at approximately 0 006 inch It is apparent from this curve that no limit had been reached for absolute accuracy of interpolation in this experiment The excellence

sions, although the method of measuring response time may have been too crude to demonstrate minor relationships that might have been present On the other hand the response times for the night viewing conditions are consistently higher than under the daylight conditions It is quite possible that this latter finding was an arti-

fact resulting from a slight delay between opening of the shutter and the fluorescing of the scale marks

Discussion

Effect of illumination on interpolation accuracy It is apparent from Table 33 1 that there was no difference in accuracy of dial reading for day and night conditions except for the dial with the most closely spaced graduations This finding is in general agreement with results obtained by Spragg and Rock (6) in an investigation of the accuracy of interpolation as a function of illumination These investigators found accuracy to be almost constant down to a scale mark brightness of 0 022 foot lambert In the present experiment the brightness of the scale markings under the simulated night conditions is estimated to have been considerably above the 0 022 foot lambert value below which Spragg and Rock found a marked loss in accuracy of interpolation

Effect of separation between scale marks on speed of reading It is noteworthy that in this experiment no relationship was found between the space between graduation marks and speed of reading, although it is admitted that the method of measuring speed of reading lacked precision In experiments conducted at Princeton University reading time increased as the space between marks was decreased but in all cases there were changes in other variables which could have caused the changes in reading speed In the first experiment by Kappauf, Smith, and Bray (3) the reduction in space between scale marks was accompanied by reductions in all other dial dimensions In the second experiment by Kappauf and Smith (4) reduction of the space between marks was accompanied in some cases by reduction in all other dimensions, in other cases by an increase in total range of values covered by the scale The question of whether or not the separation between scale marks in isolation has any

FIGURE 33 3 *Absolute error of interpolation as a function of length of graduation interval*

effect on speed of dial reading does not appear to have been answered definitely

Effect of dial dimensions on relative and absolute accuracy of interpolation In discussing the findings regarding accuracy of interpolation the distinction must be constantly kept in mind between accuracy relative to the interpolation space, and accuracy in absolute units such as degrees or inches It is apparent from Fig 33 2 that there is scarcely any useful gain in relative accuracy of interpolation as the graduation intervals are increased beyond 0 25 inch On the other hand as the intervals are reduced below this value relative accuracy falls off very rapidly Approximately one fourth (0 25) to one-half (0 50) inch would therefore seem to be an optimum value for graduation intervals from the standpoint of relative accuracy

For maximum accuracy in an absolute sense it would appear from Figure 33 3 that the optimum graduation interval, if there is such, is below the range covered by the present experiment The data in Figure 33 3 suggest that the absolute value of interpolation errors might continue to decrease with decreases in graduation interval until the limit of visual acuity is reached If this is true the limit of visual acuity would determine the optimum graduation interval for maximum accuracy of dial reading The data of Loucks (5) and Kappauf, Smith, and Bray (3) suggest that as the distance between graduation marks is decreased, there is an increasing tendency to make comprehension errors, that is, assign the wrong values to scale marks Also, Kappauf and Smith (4) have found that increasing the total number of marks on the scale increases the time required for reading It would seem, therefore, that there is no easy answer to the problem of what is the optimum interval size for instrument scales, but that the optimum interval will vary with reading criteria

Summary

Measurements were made of the accuracy of interpolating pointer position between scale marks as a function of dial diameter and the angular spacing between divisions Subjects were required to estimate the pointer position to within one tenth the space between graduations The experimental dials were painted with yellow fluorescing paint on a black background and were read under simulated daylight (45 foot candles) and night (ultra violet) illumination conditions The major results of this investigation may be summarized as follows

1 Dial diameter and angular spacing of the scale marks could be combined into the single variable of length of graduation interval

2 The relative error of interpolation decreased as the length of the graduation interval increased up to approximately 0 5 inch, and was very nearly constant at higher intervals (see Figure 33 2)

3 The absolute error of interpolation increased very nearly as a linear function of the length of the graduation interval (see Figure 33 3) If there is an optimum interval for absolute accuracy it would appear to be below the interval lengths used in this study

4 Except in the case of the most closely spaced divisions the accuracy of interpolation was independent of the two illumination conditions

5 The speed of dial reading was not systematically related to either dial diameter or angular spacing of the divisions, although the measurements were admittedly crude Slower reading under the simulated night (ultra violet) lighting conditions was probably due in part to delay in fluorescence of the dial markings

References

1 Grether, W F, Factors in the Design of Clock Dials which Affect Speed and Accuracy of Readings in the 2400 Hour Time System, *Journal of Applied Psychology* 1948, Vol 32, 159–169

2 Grether, W F, 'Psychological Factors in Instrument Reading I The Design of Long scale Indicators for Speed and Accuracy of Quantitative Readings,' *Journal of Applied Psychology* 1949, Vol 33, 363–372

3 Kappauf, W E, Smith, W M and Bray, C W, Design of Instrument Dials for Maximum Legibility I Development of Methodology and some Preliminary Results," USAF Air Ma

teriel Command Memorandum Report No TSEAA 694 1L, 20 October 1947
4 Kappauf W E, and Smith W M, Design of Instrument Dials for Maximum Legibility II A Preliminary Experiment on Dial Size and Graduation, USAF Air Materiel Command Memorandum Report No MCREXD 694 1N, 12 July 1948
5 Loucks, R B Legibility of Aircraft Instrument Dials The Relative Legibility of Tachometer Dials AAF School of Aviation Medicine, Project No 265 Report No 1 1944
6 Spragg S D S, and Rock, M L Dial Reading Performance as Related to Illumination Variables I Intensity USAF Air Materiel Command Memorandum Report No MCREXD 694 21 1 October 1948

The Accuracy of Precision Instrument Measurement in Industrial Inspection *

C H LAWSHE, JR, and JOSEPH TIFFIN

The authors acknowledge the assistance of Mr O D Lascoe in establishing the true dimensions and of Mr R N Purcell in doing most of the statistical work in connection with the second part of the study

Modern industrial production is becoming more and more dependent upon the accuracy of precision instrument inspection Thousands of employees have been trained in the use of precision measuring instruments, and future industrial developments almost certainly will require still greater emphasis on the accuracy of measurement to insure production which satisfies the fine tolerances of modern precision equipment Virtually every precision instrument calls upon the operator to exercise judgment in determining proper feel," tension," "drag,' or other characteristics In spite of all that is known about the variability of human judgments, little attention has been given to the importance of such variability as it may affect the accuracy of precision instrument measurement The purpose of the investigation reported in the present paper was to examine the accuracy and variability of employee measurements with certain precision instruments

Sources of data Data were collected in two different plants The first of these is engaged in the manufacture of variable pitch propellers for aircraft and the second in the manufacture of precision parts for aircraft and automobile engines There is no evidence that the survey results obtained are any better or any worse than those which would be obtained in other plants of a similar character and there is every reason to believe that similar results would be obtained if the survey were projected to other plants

An Inspection Department Survey

Job analysis Approximately 200 people are employed in the inspection department of the first plant Their jobs were analyzed by job classifications in order to determine what precision instruments are used and what tolerances are demanded on each job. Frequency counts were then made to determine which instruments or combinations of instruments are used in the largest number of classifications and by the largest number of employees On the basis of this count, 20 instruments and combinations were chosen as being most important in this particular plant

A dimensional control laboratory A room was set aside as a dimensional control laboratory and 20 booths or inspection stations were set up Each booth was num

* Reprinted from *Journal of Applied Psychology* Vol 29, No 6, December 1945

bered and in it were placed one of the 20 instruments, a standard part from the plant, and a simplified working drawing which indicated one dimension to be meas ured with the instrument provided When an employee entered the room, the attendant determined his job classification and provided him with an appropriate work sheet for each of the stations containing work samples from his job Each employee was tested on only those instruments which he uses on his particular job He was en couraged to make 5 measurements and then to record his best judgment as to the dimension The readings thus obtained were compared with so called true' di mensions which were determined by means of ultra precision instruments in combina tion with Johansen blocks Instruments utilized in the performance testing were checked and adjusted periodically to insure constancy

Emphasis of testing This performance testing procedure was organized in con nection with a training program and its primary function was to identify persons in need of training Plans for a mainte nance program were also made with pro visions for re testing employees every 3 months It was also planned to utilize the laboratory to supplement seniority in de termining adequacy in connection with transfers and promotions The program was instituted with the knowledge and backing of line supervision and of the union in the plant There is every reason to believe that nearly all of the employees approached the test situation with a favor able attitude

Results

Results obtained at 11 of the 20 stations are presented in Figure 34 1 The partic ular stations selected for illustration were chosen in terms of general familiarity with the instruments used and not because of any peculiarity in the findings, they are truly representative

In Figure 34 1, the open bars represent the percentage of inspectors tested who obtained readings within the established tolerance The solid bars represent the percentage of persons tested who failed to meet the standard As already stated, not all of the inspectors were tested at each station, instead the sample contains only those who use the instruments on their jobs This accounts for the fact that the N s range from 117 to 162 The figure in dicates that the percentage of inspectors meeting the various standards ranged from a high of 66 per cent on the inside mi crometer to a low of 9 per cent on the

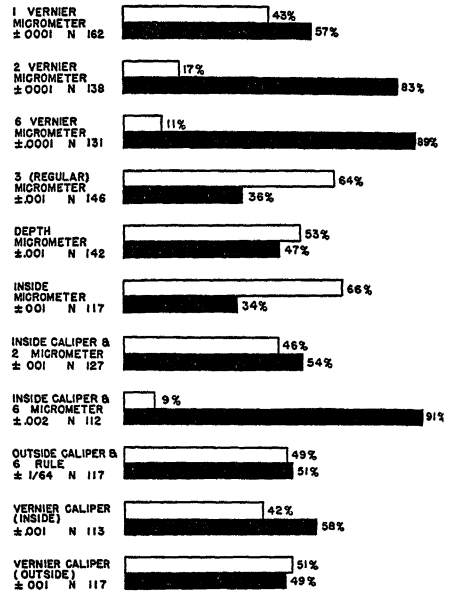



FIGURE 34 1 *The percentage of inspectors passing and failing various precision meas uring instrument performance tests in an aircraft propeller plant The open bars indi cate the percentage meeting the standard and the solid bars indicate the percentage failing*

inside caliper in combination with the 6 inch micrometer The pattern of perform ance on the various vernier micrometers also seems significant It will be noted that 43 per cent of those tested met the stand ard with the 1-inch micrometer, 17 per cent with the 2 inch, and only 11 per cent with the 6 inch The varying tolerances established for the instruments are the same as the tolerances which job analyses indicated had been established by the engi neering department and are identical with those encountered in the shop

A Tool Room Survey

Procedure Because many of the em ployees ın the plant just described were new and were drawn from a tight labor market, a related study dealing with ex perienced toolmakers was set up ın a plant engaged ın the manufacture of precision parts for aircraft and automobile engines In this study, 45 men were selected from the tool room Their ages ranged from 17 to 62 years, their experience with the com pany from 5 to 29 months, and their ex perience on their present jobs from 1 to 29 months For the most part, the job classi fications of these men fall ın higher labor grades than do those reported ın the ın spection department study

The study was limited to the use of vernier micrometers and employed 19 parts to be measured Five parts were cylındri cal, 5 rectangular, 5 spherical, and 4 were inside diameters Here again, each em ployee measured each part independently 5 times and the reading recorded was the best judgment he could make as to the true reading on the basis of these trials After all readings had been completed by all of the men, the parts were measured with ultra precision instruments and Jo hansen blocks ın order to obtain the closest possible approximations to the true di mensions against which to compare the measurements made by the men

Accuracy results No significant correla tions were found between accuracy of measurement and either age, length of ex perience with the company, or amount of time on the present job The percentages of accuracy are shown ın Figure 34 2 On the baseline of this figure ıs plotted the approximate size of the part On the ver tical axis ıs plotted the percentage of read ings correct to 0001 inch It will be noted that the parts vary ın size from approxi mately ¼ inch to approximately 6½ inches Each of the four lines plotted ın Figure 34 2 shows the percentage of read ings within 0001 inch of the ' true dimen sions for parts of a certain shape according to the code indicated ın the figure Thus, for cylindrical parts, about ¼ inch ın size, 73 per cent of the readings were accurate to 0001 inch For cylindrical parts approx- imately 3 inches ın size, however, only 12 per cent of the readings were accurate to 0001 inch

Variability results The results plotted ın Figure 34 2 assume the accuracy of the so called true dimension The validity of these true dimensions ıs always open to

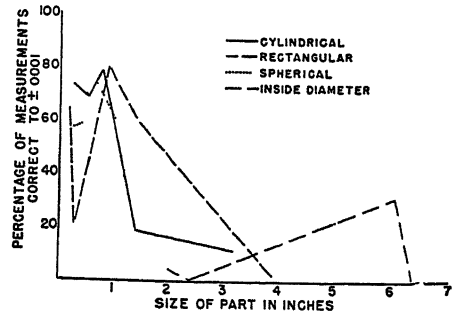



FIGURE 34 2 *The percentage of toolmakers (N 45) who obtained vernier micrometer measurements within 0001 inch of the true dimension on each of nineteen parts The base line indicates the approximate size of the dimension and the code indicates the shape of the part measured*

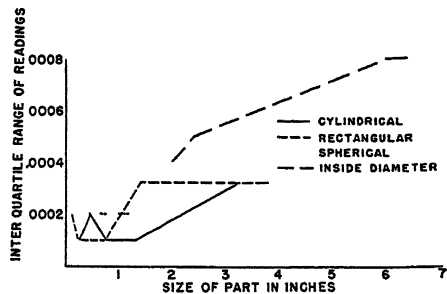



FIGURE 34 3 *The variability of the vernier micrometer measurement of 45 toolmakers on each of nineteen parts The base line indicates the approximate size of the dimen sion and the code indicates the shape of the part measured*

question ın spite of the ultra precision methods used Therefore the data were analyzed ın another way to show the vari ability of the readings obtained without reference to the true' dimensions This analysis of the results ıs plotted ın Figure 34 3 Here again the approximate size of the part ın inches ıs plotted on the baseline

The vertical axis on this chart plots the interquartile range of the readings This may be interpreted to mean the range of the middle 50 per cent of the readings under each condition of size and shape of the part For example, looking at the solid line plotted in Figure 34 3, it will be noted that when the part is cylindrical and approximately ½ inch in size the interquartile range of readings, or the middle 50 per cent, is 0002 inch This means that 50 per cent of the readings were within a range of 0002 whereas the remaining 50 per cent of the readings varied by more than 0002, either above or below the range of the middle 50 per cent In like fashion, it will be noted that when the part is cylindrical and approximately 2¼ inches in size, the middle 50 per cent of the readings fall within a range of 0003, whereas the remaining 50 per cent of the readings on this part fall more than 0003, one way or the other, from this range

Summary and Conclusions

Controlled performance tests with various precision instruments were administered in two industrial plants In one plant 200 inspectors were tested on a variety of instruments used in their respective jobs In the other, 45 tool room employees were tested on vernier micrometers with parts of varying sizes and shapes

In general, the following conclusions are supported

1 The accuracy of precision instrument usage is probably considerably lower than is ordinarily assumed by those responsible for methods and standards In one plant the percentage of inspectors meeting the standard ranged from 66 per cent on the use of the inside micrometer to 9 per cent on the inside caliper in combination with the 6 inch micrometer

2 In the population studied, micrometer reading accuracy did not correlate significantly with age, amount of experience with the company, or length of time on the present job

3 Gross size of the part is apparently a factor in the accuracy of micrometer measurement Under optimal conditions in the second plant, not more than 80 per cent of the readings were accurate to 0001 inch and with larger parts, ranging from 3 to 6 inches, only about 20 per cent of the readings were accurate within these limits

4 Gross size of the dimension is likewise related to the variability of measurements As the parts increase in size, regardless of shape the spread of the readings becomes greater so that for large inside diameters, 50 per cent of the readings vary by 0008 inch from the other 50 per cent of the readings

5 The results suggest that while inside diameter measurements are more variable than measurements of cylindrical, rectangular, or spherical dimensions, the percentage of measurements meeting the standard of ± 0001 inch is no less However, the problem of the relationship between shape and both accuracy and variability is open to further investigation

6 The necessity for the development of training methods that will more nearly standardize judgments based on such characteristics as feel,' 'tension,' and drag' in the use of precision instruments is indicated

7 The implication is present that the very nature of the vernier micrometer and similar precision measuring instruments is such that one should not expect as high a degree of constancy as the average operator, supervisor, and standards man has been taught to expect

Auditory Signals for Instrument Flying *

T W FORBES

Introduction

In 1936, de Florez demonstrated that a pilot could fly an airplane with his eyes blindfolded when two of his instrument indications were given by means of an auditory signal in his earphones The pilot flew with rudder and elevators only, allowing the inherent stability of the Fairchild 24 to take care of lateral control The signals employed were (1) a turn indication consisting of increase of the signal intensity to one ear and decrease of intensity to the other and (2) an air speed indication consisting of a change of pitch of the signal Although the plane described a wide climbing spiral rather than a straight path, the pilot was able to maintain satisfactory control and to recover from spins, thus demonstrating the feasibility of flying by auditory indications

It is well known that the large number of operations required of the pilot of a modern airplane, and the multitudinous instruments which he must follow, tax the abilities of the pilot to the limit His eyes are especially subject to overload, since he must keep track of the large number of instruments on the panel and at the same time look out for obstacles and observe his position relative to the ground In fact, in the larger airplanes both pilot and copilot are sometimes kept busy, especially when flying under 'instrument conditions

The continuously increasing speed of modern planes and the advent of new devices such as air borne radar will, if anything, increase this critical load on the eyes Thus it would be an advantage if some of the flight indications could be furnished to the pilot through the ears Accordingly, a further study [1] of auditory flight indications has been carried out during the last 2 years, the results of which will be reported in this article

The purposes of the study were to determine (1) what types of auditory signals could be followed with greatest ease, (2) with what accuracy such signals could be utilized, and (3) how many simultaneous auditory signals could be followed successfully It was therefore necessary, first, to design a variety of different auditory signals and, secondly, to test their effectiveness on a group of men in a task similar to that of 'flying blind' in an airplane

The results indicate that, if the signals are properly designed as many as four auditory indications can be followed without interfering with radio and interphone communication when occasion demands

Development of Successful Auditory Signals

Tone signals The aim of this part of the investigation was to develop three aural indications that could be followed simultaneously Indications for turn, bank, and air speed were tried out on two synthetic devices One of these was the airplane pursuitmeter, a device in which the subject tried, with controls similar to airplane controls, to compensate for deviations introduced in unpredictable fashion by a

* Reprinted from the *Journal of the Aeronautical Sciences* Vol 13, No 5 May 1946

[1] At the Psycho Acoustic Laboratory, Harvard University The research was begun under contract with the Office of Scientific Research and Development, at the request of the U S Navy, Special Devices Division, and is being continued under a contract with the U S Navy, Office of Research and Inventions It is desired to acknowledge indebtedness to Rear Adm Luis de Florez, U S Navy Office of Research and Inventions, for making the work possible to Mr W R Garner and Miss Jean Howard for assistance with the experimental work and to the members of the Civilian Public Service Unit assigned to this laboratory for serving as subjects in a long series of tests

mechanical device Automatically com puted scores measured the success with which the signals were followed The second device used was the familiar Link Trainer In this case the auditory signals were arranged to give turn, bank, and air-speed indications In both cases the records obtained by use of the auditory signals were compared with records made by the same individuals using visual indications alone (without any compass or gyro direction indicator) and with the rough air attachment turned on The task set by means of the airplane pursuitmeter was a somewhat similar one A number of trained pilots also tried the best signals on the Link Trainer

Several types of auditory signals were tested, the first being combinations of tones, pitches and chopped signals In general, it was found that those combinations that

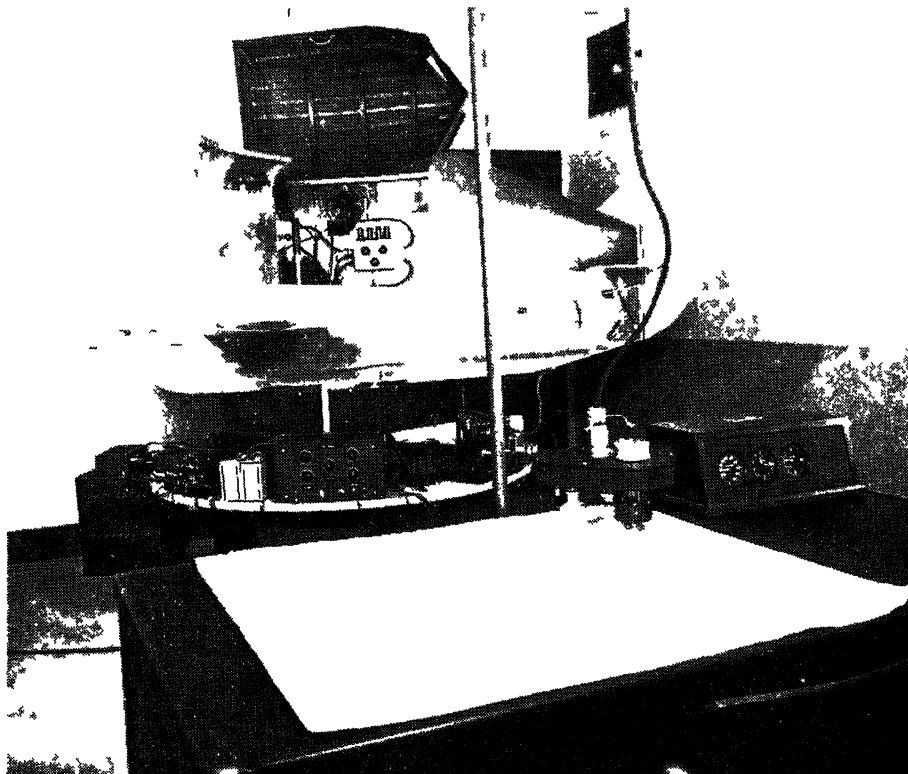



FIGURE 35 1 *The Link Trainer equipped for testing auditory signals Signal generators and servo units are shown mounted around the base of the trainer The usual crab recorder appears on the desk in the foreground*

were formed by adding separate tone sig nals together were too complex to be fol lowed successfully Some individuals had difficulty in analyzing each signal from the complex auditory pattern when two or more signals were used There was a gen eral tendency, when using these complex signal combinations, for the subject to fol low one indication with such attention that

Ten men between the ages of 19 and 36 were used as subjects Since none of the subjects had had any previous flying train ing, the course of learning could be fol lowed for both the visual and auditory types of indication

The task in the case of the Link Trainer, was to fly a straight course by means of turn, bank, and air speed indi

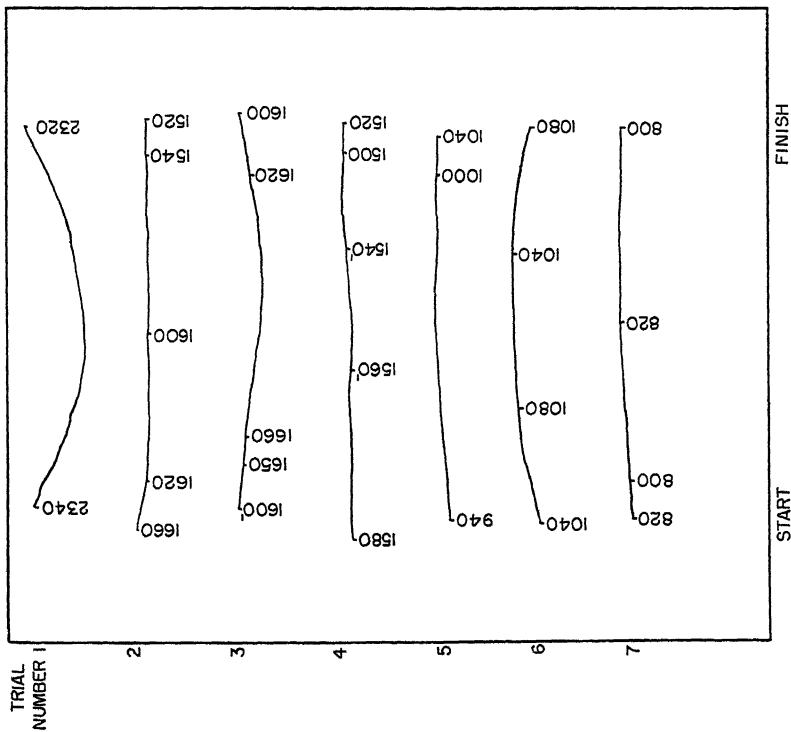



FIGURE 35 2 *Learning series using visual indications Link Trainer records for subject C Trials were 10 min in length*

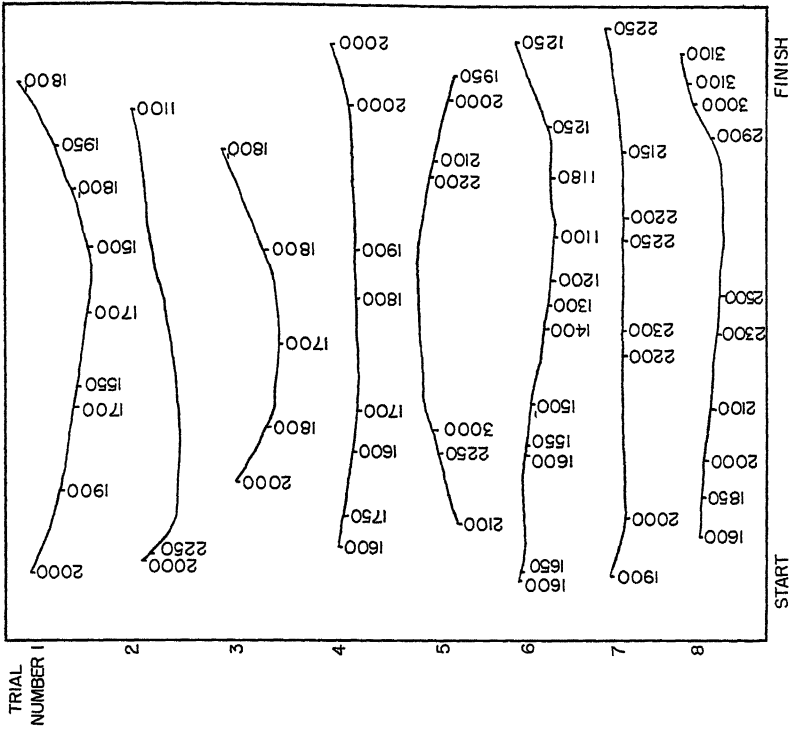



FIGURE 35 3 *Learning series using three in one auditory indications Link Trainer records of subject C who was representative of the better performers Trials were 15 min in length*

the other two escaped him and went out of control

However, two auditory signal combinations were developed which were successful enough to be of possible practical utility Both of these sounded like the behavior of the airplane One of them was similar to a signal used by de Florez It involved apparent motion of the tone to the right or to the left in a manner roughly corresponding to the right and left indications of a radio compass This signal can be used with a second indication superimposed on it to indicate air speed

The second successful signal was a 3 in-1 indication That is to say, different characteristics of the same signal were used to indicate turn, bank, and air speed These indications were (1) a repetitive or sweeping type of motion of the signal from left to right, (2) apparent tilt' produced by pitch variations, and (3) a "putt" sound the rate of occurrence of which could be associated with the sound of the airplane motor

With this 3 in 1 signal the 10 subjects learned to operate the Link Trainer so as to fly a respectably straight course in "rough air" after a couple of hours of practice The results corresponded quite well with those obtained by the same men using visual indications after about the same period of practice Fig 35 1 shows the signal equipment mounted on the Link Trainer for test purposes, while Figs 35 2 and 35 3 give samples of records made with both types of indication As would be expected, some of the men had more difficulty than others, both in learning to manipulate the controls and in following each type of signal

Six private pilots, some of whom had had some instrument instruction, were able to operate the Link Trainer creditably by means of the auditory signals after a practice period of about 1 hour Several navy pilots, after a short trial, were confident that it would be possible to fly with this signal combination

The 2 in 1 signal did not give so accurate results as the 3 in 1 but could be used successfully where two signals were sufficient

Either of the two signal combinations might be applied to other instruments than the ones used for test purposes For instance, the turn signal might be of more use if controlled by the gyro direction indicator or the radio compass for keeping on course during long flights

It was of especial interest that the simulated radio range signal and voice communication could be heard simultaneously with the 2-in 1 and the 3 in 1 type of signal without difficulty This was due to avoidance of a large number of tones, which would cause greater interference with range and voice signals

Automatically produced speech signals A device called an automatic annunciator was developed for the purpose of announcing altitude, air speed, or other similar instrument indications directly to the pilot This device translated the indicator readings automatically into spoken messages It proved successful and offered promise of useful application for many types of instrument indications and of warnings

As developed for demonstration purposes, the device announced altitude in 200 ft units through the pilot s earphones The annunciator consisted of a light, compact, multichannel sound reproducer of the magnetic tape variety, on each channel of which a permanent message had been previously recorded Each channel contained one spoken number or one unit of the message, and these were selected by relays operated by a control switch unit This unit followed the Link Trainer altimeter by means of a self synchronous repeater and servo units After the appropriate message units were selected, they were played in the proper sequence in order to say, for example two thousand two hundred feet

The automatic annunciator may be coupled to instruments by servo devices in cases where the aircraft instrument does not have sufficient torque to operate the control device direct However for some applications, the annunciator may be operated direct (without any servo unit) by means of vacuum tubes

The control circuit is such that, with a 4 unit message such as four thousand four hundred feet, the lag in the spoken indication will never be more than the

length of the message plus 1 sec In the 4 unit message this means a lag of 5 sec In case the application requires the shortest possible time interval, single channel messages may be used, with a resultant lag of about 1 sec, and, in case this 1 sec lag is still too great, a circuit of the anticipation type can be employed to reduce it still further

FIGURE 35 4 *The multichannel magnetic tape sound reproducer The magnetic tapes carrying spoken message units are carried on the edges of aluminum discs which form the reproducer drum This drum rotates beneath pickup coils which feed the signal to an audio amplifier*

The 24 channel magnetic tape annunciator[2] weighed approximately 15 lbs Associated selection and control equipment will weigh more or less depending on the application Fig 35 4 shows the multichannel sound reproducer, and Fig 35 5 illustrates it covered and connected to a relay and sequence unit developed for use with the altimeter

It is of interest that trials of the annunciator on the Link Trainer altimeter have shown that there is little difficulty in distinguishing between the speech from the annunciator and incoming speech communication from outside (radio) channels This is due to the fact that communication messages may be made to override the annunciator and also to the fact that the annunciator has a repetitive rhythm and a speech characteristic which are readily identified

Possible applications It is not anticipated that auditory signals will entirely displace visual instruments in the airplane for any of the fundamental indications However, it has been demonstrated that as many as three auditory signals can be followed with accuracy if properly designed aural indications are used It is thought

[2] Developed through the cooperation of Bell Telephone Laboratories Appreciation is expressed to Dr E C Wente of N D R C Section 17 3 and Mr W L Woolf of Stevens Institute of Technology for valuable advice and interest

that the spoken indications can be useful as auxiliary aids to furnish warnings of various types and to give altitude and air speed information to the pilot at times when his eyes must be otherwise engaged

A tone signal attached to the gyro direction indicator might be of considerable advantage on airplanes that do not have the automatic pilot for maintaining the proper heading on cross country flights Or again such a signal might be controlled by the self orienting loop of a radio compass for long distance homing' on radio stations This would allow the pilot to maintain the proper bearing on the station without the necessity of constantly watching the needle indicator of the radio com pass

It is possible that auditory signals, if properly designed, can be of assistance in connection with some of the new blind landing systems that are under develop ment

The automatic annunciator can be arranged to announce altitude or air speed, or both alternately, for landings and take-offs Such auditory messages would be of advantage in single seater planes for land ings made under low visibility and night conditions On larger airplanes they could also relieve the copilot of calling out this information Furthermore, the annunciator can be arranged to call attention to sources of trouble, such as wheels not down, flaps not down, gas tank empty, and other items

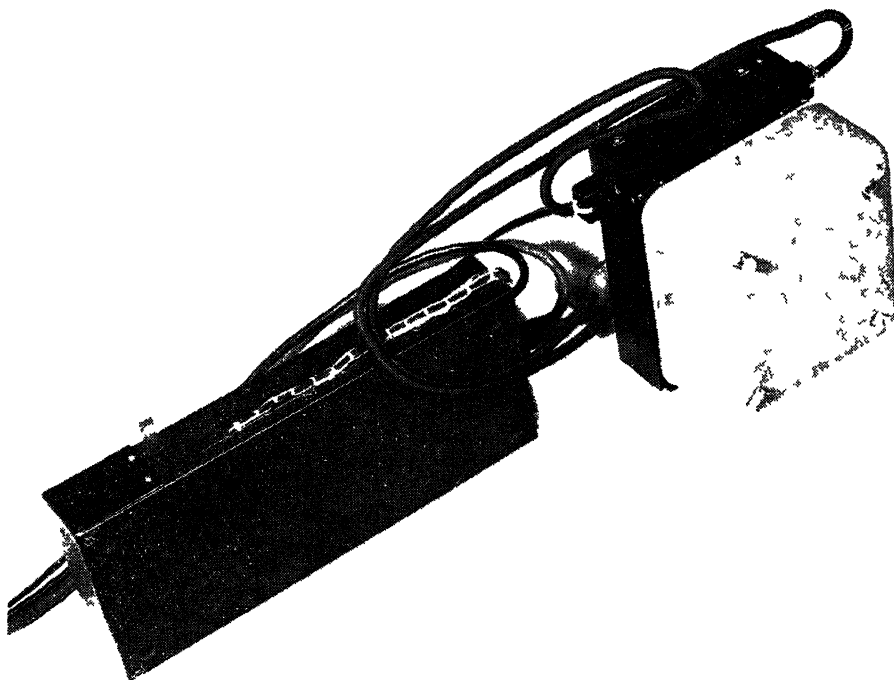



FIGURE 35 5 *The multichannel reproducer connected to the relay and sequence unit This latter unit was developed for announcing altimeter readings*

Fundamental Principles of Successful Aural Indications

In order to be successful auditory signals must be designed with a view to certain psychological principles having to do with hearing and with the pilot s reactions Cer tain of these principles may be summed up as follows

(1) Pilots have certain habitual methods

of thinking about the attitude of the air plane and similar variables The signals must be designed to fit these habits of thought

(2) Most fliers are much more accus tomed to using visual indications than au ditory ones Therefore the auditory signals should be as simple and as self explanatory as possible Only in this way will the need for special training of pilots in use of the signals be reduced to a minimum

(3) Under most circumstances, a person can attend to only one thing at a time It was found that, when several independent auditory signals were added together, the individual following the signals might attend for a while to each in turn, as he was supposed to do Then however, there was a tendency for one signal to capture' his attention to the exclusion of the other To be of practical value, the auditory signals must be designed so that this capture of attention does not occur

(4) In military aircraft, especially, in formation from various sources is contin uously obtained through the earphones from radio and interphone channels Any auditory indications used must be designed to interfere as little as possible with such messages

(5) It is not sufficient to produce an auditory signal that satisfies the designer The signals must also be tested bv psycho logical methods on persons unbiased by familiarity with their development They must be made to fit the capabilities of the average pilot, not merely the special au ditory talents of a few individuals

It is believed that auditory signals de veloped in accordance with these principles offer considerable promise for the future of aviation Auditory signals have a distinct advantage wherever the pilot s attention must be called to a malfunctioning part or whenever his vision must be otherwise occupied as in takeoffs and landings

REFERENCES

1 de Florez Luis, True Blind Flight *Journal of the Aeronautical Sciences* Vol 3, 1936, pp 168–170

Chapter IX

DESIGN OF CONTROLS

Control design is primarily concerned with the most efficient utilization of human effort in directing a machine As technological development enables more energy to be available at man's finger tips it becomes increasingly important that precise control of this energy be exercised

Ordinarily displays and controls are regarded as separate entities, but this dichotomy is not always distinct For example, the gear shift lever on a car acts not only to manipulate the shifting of gears and therefore as a control of energy, but also to transmit information to the driver by its position and is therefore a display Certain airplane controls also act simultaneously as kinesthetic displays

The study by W O Jenkins on "The Tactual Discrimination of Shapes for Coding Aircraft Type Controls" is a case in point Fitts regards knobs as tactual displays and yet they are also controls Display or control, the design of distinctive knobs is important in the operation of machines Jenkins' study determines the shape of knobs revealing a minimum number of intra set errors

A driver might be aided considerably if knobs for cigarette lighter and parking lights, among others, could be readily discriminable especially under night driving conditions when touch rather than vision is the cue

Control knobs are often taken for granted, and such variables as ratio of knob-to-pointer movement, knob diameter, use of a crank handle, or effect of backlash

are often overlooked or neglected The research of Jenkins and Connor establishes an optimal ratio of one or two inches of pointer movement for one complete turn of the knob It also establishes that this factor is of greater importance than knob diameter, crank handle, or backlash in the situation investigated

The article by Orlansky presents a technical review of the psychological aspects of stick and rudder control It is concerned with human physical requirements in relation to the psychological aspect of stick feel or handling quality The difference in force exerted in either push or pull when coupled with the problem of left- and right handedness indicates that design must take human strength limits into consideration The elevator, aileron, and rudder determine the rudiments of flying, but their manipulation and regulation are human problems of physical strength and psychological reactions The "feel" is more important than instruments under the exigencies of combat and therefore the psychological aspects of handling are of paramount importance This article by Orlansky shows the value of summarizing and evaluating pertinent writings on the subject It clears the air for further experimental work

Coakley's article establishes that human behavior alters the operation of highly automatic machines In other words, the operators of automatic equipment influence the nature of the machine product Were it not for the clear and precise manner in which Coakley presents the evidence this point would otherwise be most mystical Although this article is not too appropriate to this chapter, it is both too important and too good to be omitted from this collection of readings

The Tactual Discrimination of Shapes for Coding Aircraft-Type Controls *

WILLIAM O JENKINS

Introduction

The present studies were designed to determine the extent of confusion among several series of control knob shapes, including shapes now in use in aircraft and a number of experimental ones, in a situation where the subjects employed primarily tactual cues, and the use of vision, kinesthesis, size, and position was minimized or eliminated On many occasions aircrew members, particularly pilots, must react rapidly and accurately to one of a closely bunched group of controls where positional cues are minimized and the operator is attending to certain instrument indications Errors in this regard have produced many accidents, particularly in transition training where unfamiliar aircraft are being flown One of the most common accidents results from confusion between flaps and landing gear controls which are in close proximity in some aircraft and are not coded with respect to any of the several possibilities

Investigations by Weitz at the Army Air Forces School of Aviation Medicine have shown that accuracy of performance is significantly affected by position cues and by shape and color coding of aircraft type control knobs The present studies provide specific information, without particular

* This chapter is based upon research findings reported in Headquarters AMC, Engineering Division, Memorandum Reports, No TSEAA 694 4, TSEAA-694 4A, and TSEAA 694 4B

samples of knob shapes, concerning which shapes yield the fewest recognition errors where size, position, vision, and mode of operation of the control are held constant or eliminated as cues

Apparatus and Procedure

In the first major study (study I), 25 plastic shapes, 1½ inches in the largest di mension, were each mounted on a rod which was bolted to the periphery of a turntable The knobs and their mode of presentation are shown in Figure 36 1 In the second major study (study II), 22 plastic shapes, 1¼ inches in the largest di mension, were mounted on the turntable with their shafts parallel to the ground as shown in Figure 36 2 In both studies the procedure was followed of examining standard control knob shapes in the cock pit and on AAF equipment such as radar and radio sets and attempting to select shapes for study which maximized the characteristics typical of a group of related knob shapes

In these studies the practical interest was in finding sets of knob shapes yielding a minimum number of errors for use in the cockpit and on radio and radar equip ment For this reason separate studies were made of different series of knob shapes mounted vertically and horizontally It is quite possible that mode of mounting is not a critical factor, evidence reported below suggests it is not

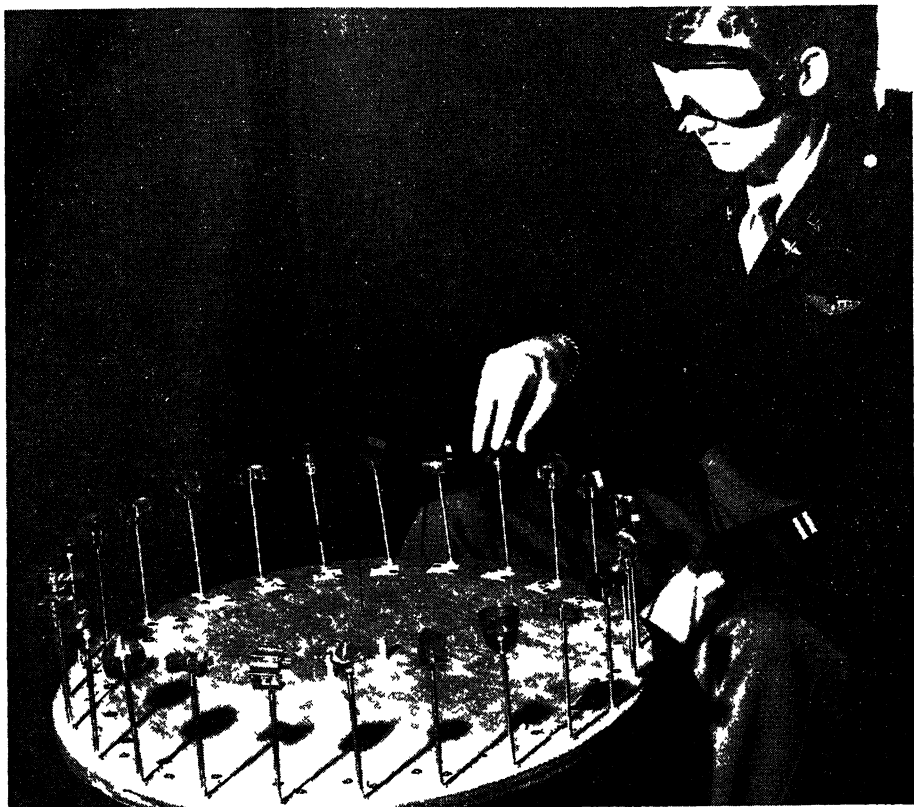



FIGURE 36 1 *Vertically mounted knobs employed in a study of shapes for use in coding aircraft control knobs*

The procedure employed was as follows Each subject was seated before the turn-

table and the instructions were read to him A given knob was then presented to the blindfolded subject who felt it for 1 second The experimenter then rotated the turntable to a pre designated point and the subject went from knob to knob feel ing each one in turn until he found and reported what he thought was the test jects started their test under one condition and the other half under the other In ad dition, the order in which the test and comparison knobs were presented was varied systematically in order to check in traserial effects

In both investigations two types of errors were recorded The first was the obvious

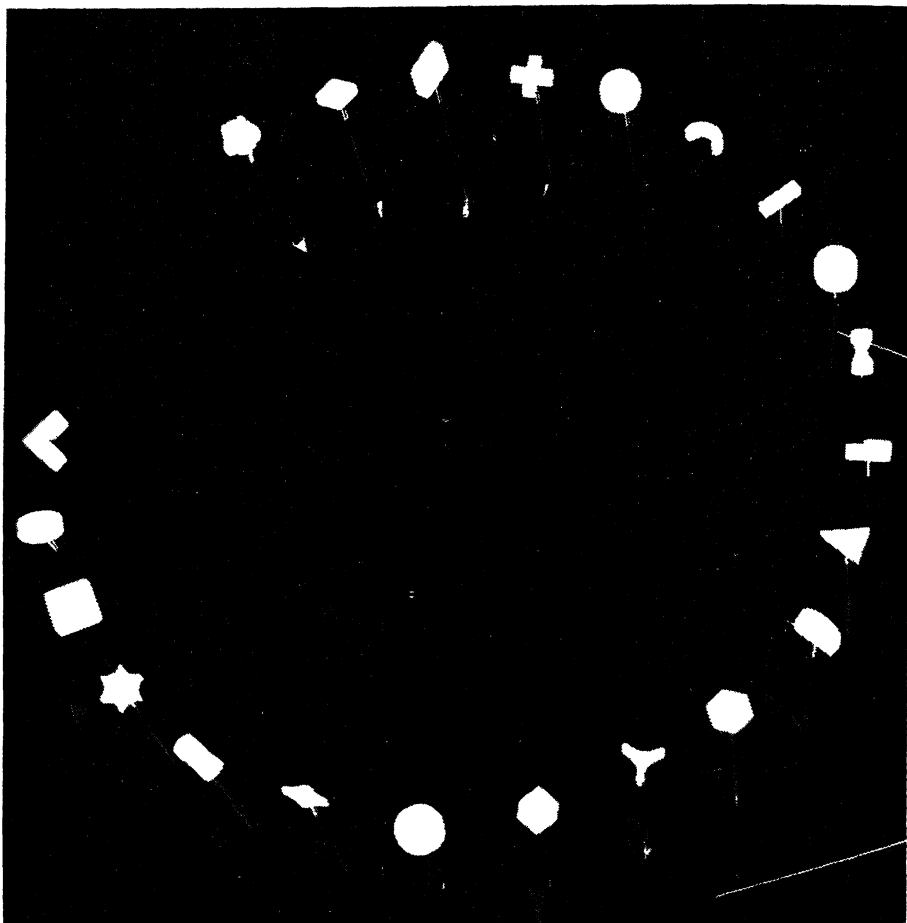



FIGURE 36 2 *Horizontally mounted knobs employed in a study of shapes for use in aircraft control knobs*

shape The same procedure was followed for each of the knobs, once while the sub- ject used his bare hand and once while he wore a medium weight flying glove (A-11 A)

The conditions of bare hand and glove were counterbalanced so that half the sub case in which a subject incorrectly identi fied one shape as another The second was called a hesitation error and was defined as the case in which a subject spent more than the allotted second in handling an incorrect shape, but did not identify that shape as the correct one This type of

pause might well be undesirable in the operation of controls in the cockpit

A paired comparisons follow up study was conducted using the eight best shapes of the first investigation This investigation is described in a later section

The subjects in these investigations consisted of two separate samples of 40 Army Air Forces pilots who appeared to be representative of the pilot population

RESULTS

It was found that the order in which the test and comparison shapes were presented did not affect the number and pattern of errors appreciably, so that the data for the several different conditions have been combined in the succeeding analyses The findings have been divided into two sections those concerned with distribution statistics, and those dealing with the pattern of errors

TABLE 36 1

Distribution Statistics Concerning the Errors Made by 40 Pilots in Two Series of Knob Shapes Employed in Investigation of Shapes for Coding Aircraft type Controls

Condition	*Study I*			*Study II*		
	Mean number of errors	SD	Per cent error	Mean number of errors	SD	Per cent error
Error, bare hand	2 9	2 2	12	3 7	2 6	17
Error, glove	4 8	2 7	19	4 9	2 8	22
Hesitation, bare hand	7 2	4 1	29	1 5	1 5	7
Hesitation, glove	9 7	6 3	40	2 1	2 1	10

DISTRIBUTION STATISTICS

The distribution statistics for the two studies are summarized in Table 36 1 It can be seen from this table that the percentage of error ranges from 12 per cent to 22 per cent and the corresponding figures for hesitation type errors vary from 7 per cent to 40 per cent A satisfactory spread of scores was obtained in every case In this connection it might be noted that practically every subject made some errors of both types

The differences in Table 36 1 between performance under conditions of bare hand and while wearing the flying glove are significant in most cases at the 5 per cent level of confidence or better

In order to obtain an estimate of the reliability of the method, rank order correlation coefficients were computed between the frequency of errors and hesitations under conditions of bare hand and while wearing the flying glove The figure for errors was 0 70 while for hesitations it was 0 72 For errors and hesitations combined the value was 0 75 This degree of consistency between the performance under the different conditions indicates a satisfactory reliability for the testing technique for the present purposes

One secondary analysis was performed in regard to the frequency of errors This break down consisted of comparing the proportion of errors for pilots in study I having less than 900 hours of flying experience with that for pilots having more than 1,100 flying hours While the more experienced pilots made a slightly higher proportion of errors under both conditions of bare hand and glove than did the group with fewer flying hours, the differences were well within the range of values expected on the basis of sampling fluctuations

PATTERN OF ERRORS

It was found that the pattern of errors was comparable for errors and hesitation type errors, for conditions of bare hand and glove, and for the different orders in which the test and comparison knobs were presented so that the data for all condi

tions have been combined in the following treatment

The next step in the procedure was to rank each knob shape according to two criteria The first was the total number of hesitations plus errors made by the subjects when a particular shape was presented The second was the total number of knobs with which a given shape was confused regardless of the number of errors or hesitations involved It might be noted in passing that the rank difference correlation between these two measures was of the order of 0 10 A two way table was then constructed in which those knob shapes yielding the smallest number of errors and smallest number of knobs with which they were confused were placed side by side Further refinements in grouping were accomplished by an empirical procedure designed to obtain the largest possible number of knobs with the fewest intra confusions The resulting two way tables for studies I and II are shown in Figures 36 3 and 36 4 From this analysis 2 sets of test knobs were derived in study I in which the number of errors are minimized and one set in study II These groups are set off by heavy lines in Figures 36 3 and 36 4, and the knobs are shown in Figures 36 5 and 36 6 The best set of 8 knobs in study I yielded a total of 6 errors or one half of 1

	16	2	6	17	15	1	13	14	4	10	3	20	25	8	7	9	24	11	23	22	18	5	19	21	12
16												2			3				1	1			2		
2											1	14				1						1	1	1	
6											1						5								30
17													4			11		21				1			
15						1				18	2											11	2	23	1
1					1				2		2			28			1				1			1	1
13		1											7			9		4		1		1	1		
14	1	1	1						10								2		2		2				
4						1		3									8			2	17	3	9	4	
10		1		1	16	1	1									1				2		5	5	32	
3					1															25	2	1	1	1	6
20		29													2				6	4			1		1
25		2		1			2		1							1		11	1	1					
8	2		1			15					2										2	1		1	2
7	2	6	1									2		2		1		2				2			
9				5			14			2		1			1			16							
24			1		2				7		1										7				3
11		1		15	1		4		1		1		5	1	1	2							1		
23	3	1					2	2	2			6			1		1	2			1	1	11	1	
22		1	1		2						36	1										1	7	1	7
18				1	1	3			12		8			1		1	6			3			2	6	2
5					11	1			1	27					2			1		1	4		7	5	2
19		1	1		3		1	1	1	4	1		1				1	1	7	3	7	4		2	1
21			1	1	36				1	33								3		1	2	7	1		
12	1	1	25								12	1		7	1		1		2	3	1	1	1	1	

FIGURE 36 3 *Error pattern among 25 vertically mounted knob shapes*

per cent of the total number of errors In study II the best set of 8 knobs yielded zero errors

An examination of Figures 36 3 and 36 4 reveals that in general knobs with similar shapes tend to be confused with one another but not with knobs of different shapes There is a suggestion that the shapes tend to be grouped into families on the basis of shape, for example, shapes characterized by corners, edges, and flat surfaces (cubes, wedges, etc), with the errors occurring within a family, but not across family lines

Suggestive evidence that the 2 modes of mounting employed in these studies were not critical for the findings may be found in a comparison of the best 8 shapes of each study in Figures 36 5 and 36 6 It can be seen that 4 of the best knobs in each study are the same in shape and one is similar One of the best shapes of study II appears in set 2 of study I The remaining knobs were not used in both studies

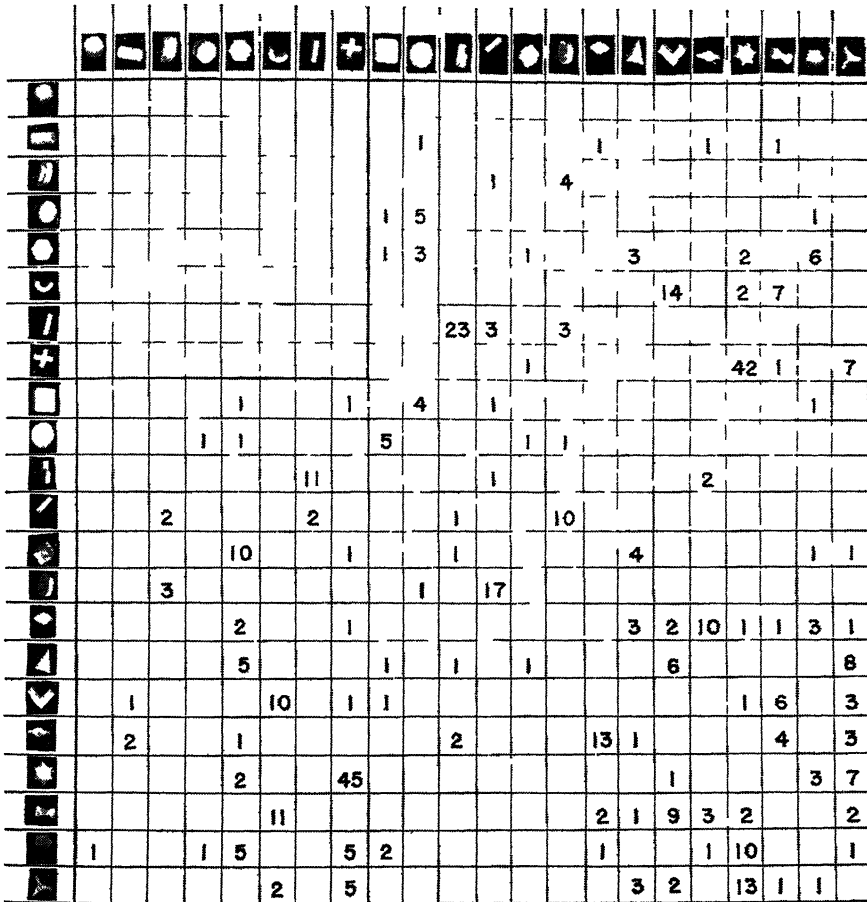



FIGURE 36 4 *Error pattern among 22 horizontally mounted knob shapes*

A follow up study was made of the best 8 knobs in study I along with 3 novel shapes employed in an ideal cockpit developed by the United States Navy Department The set up is shown in Figure 36 7 with the 3 Navy knobs in the top right hand corner The method of paired comparisons was used with the usual precau-

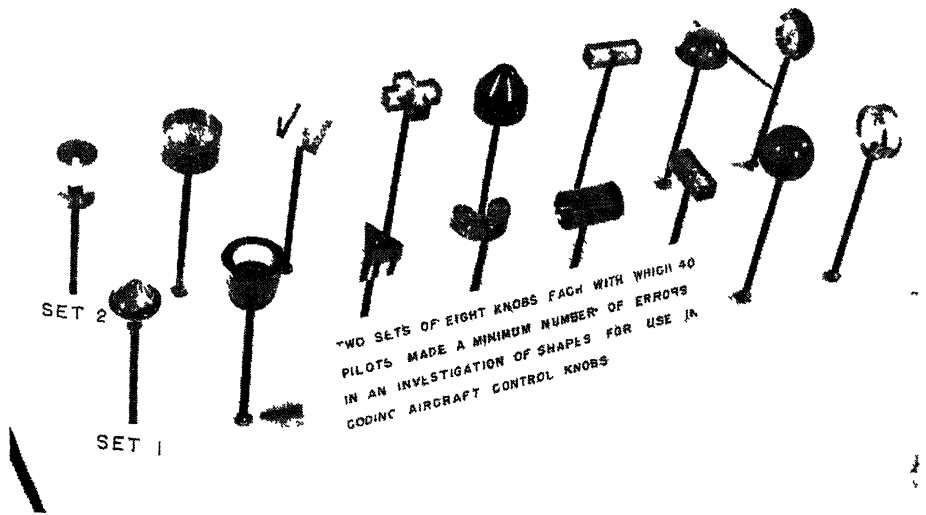



FIGURE 36 5 *Eight shapes yielding the fewest errors (set 1) and alternate shapes (set 2) in a group of vertically mounted knobs*

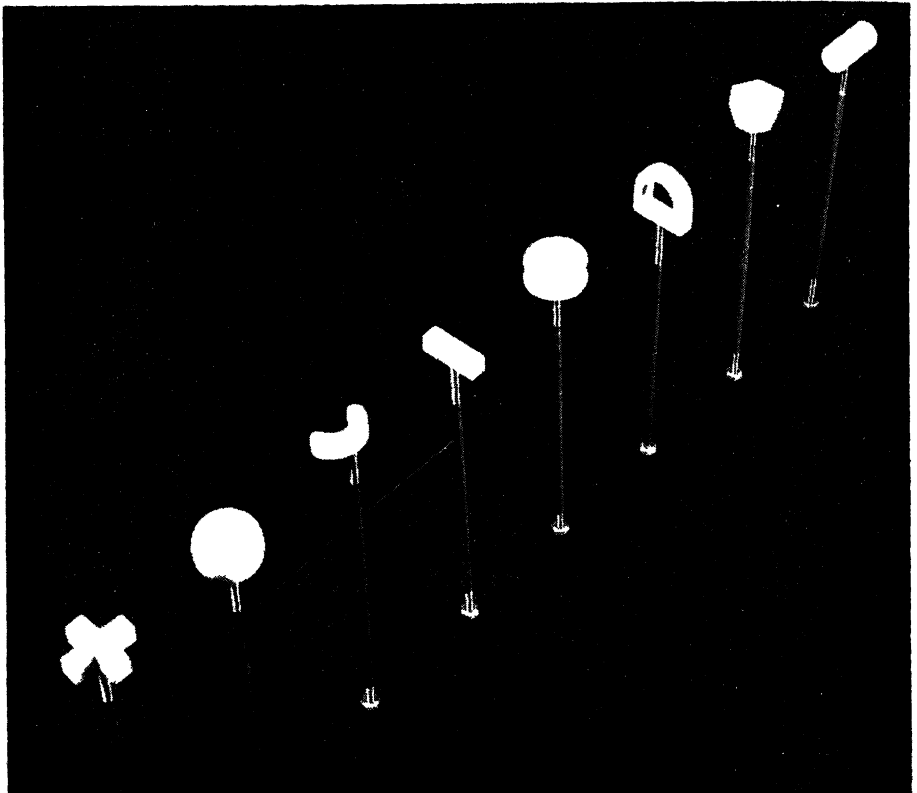



FIGURE 36 6 *Eight shapes yielding zero error in a group of 22 horizontally mounted knobs*

tions being taken of counterbalancing the order in which a given pair of knobs was presented, and varying the order of the series Each of the 11 shapes was paired with every other shape including itself once, but no knob followed itself at any place in the series A total of 30 AAF pilots was tested while wearing blacked out goggles, once with the bare hand and once while wearing a medium weight flying glove (A 11 A)

pit and on radio and radar equipment The need for standardization with regard not only to shape but also to size, position color, and mode of operation is obvious and has also been recommended

There is a need for further research in this area particularly on such problems as the optimal control handle shapes for different modes of control operation, the use of mode of operation as a cue for differential reaction, and the type of color coding (e g, brightness differentials) yielding most accurate and rapid performance under conditions of both day and night flying

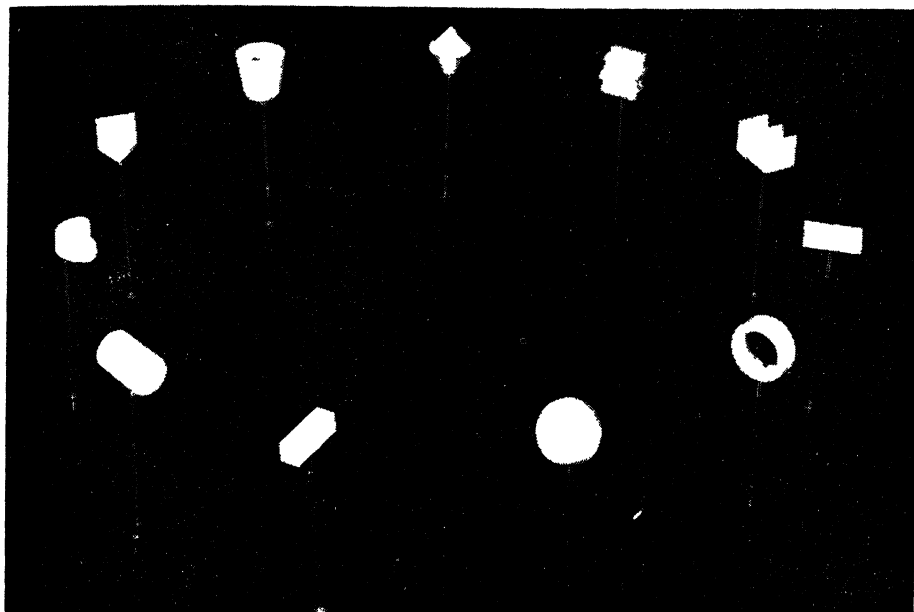



FIGURE 36 7 *Eleven vertically mounted knobs employed in a paired comparison follow up study for use in coding aircraft control knobs*

In the 1,980 comparisons a total of 9 errors was made by the 30 pilots or about one half of 1 per cent It is of interest to note that 8 of the 9 errors involved the 3 new Navy knobs which were added to the earlier set, but the N is too small for the difference to approach an acceptable level of significance

Discussion

On the basis of the present findings, recommendations have been made to the appropriate authorities that action be taken to decide which knob shapes are to be employed on which controls in the cockpit and on radio and radar equipment The need for standardization with regard not only to shape but also to size, position color, and mode of operation is obvious and has also been recommended

Summary

These studies were undertaken as a basis for selecting sets of control knobs of different shapes for use in the cockpit and on radio and radar equipment which could be recognized immediately and accurately by touch in a situation where the use of vision, size, position, and mode of operation was held constant or eliminated as a cue

Two sets of 25 vertically mounted and 22 horizontally mounted knob shapes, including some knobs now in use in aircraft and a number of experimental ones, were presented to 2 separate groups of 40 blindfolded pilots who compared each shape with every other one in the series One test was made with the bare hand and another while wearing a medium weight flying glove The findings may be summarized as follows

1 A percentage of errors ranging from 12 to 40 per cent was found Practically every pilot made several errors

2 A significantly greater number of errors was made while wearing the flying glove as contrasted with the condition of bare hand

3 The correlation between frequency of error with the bare hand and that while wearing the flying glove was 0 75, indicating a satisfactory reliability for the technique

4 Pilots with larger numbers of flying hours made slightly, but not significantly, more errors than those with fewer hours

5 Two sets of 8 knobs were found in the first study which yielded a minimum number of intra-set errors and there were no confusions among the best 8 shapes of the second study

6 A paired comparisons study of the best 8 knobs of the first study along with 3 additional knobs yielded a very small number of errors

Recommendations were made to the appropriate authorities that the shapes yielding the fewest errors in these studies be employed on aircraft equipment, and that the use of these knobs be standardized on equipment of a given kind

Some Design Factors in Making Settings on a Linear Scale *

WILLIAM LEROY JENKINS and MINNA B CONNOR

This research was executed under Contract No W28 099 ac 130 between the Institute of Research, Lehigh University, and the USAF Air Material Command, Watson Laboratories, Red Bank, N J

In setting a pointer on a linear scale by means of a control knob, is there an optional ratio between pointer movement and knob turn? Is there an optimal knob diameter? When is a crank handle better than a knob? What is the effect of backlash in the system? No previous systematic investigation of such design factors seems to have been made

The present study deals with a situation in which the operator is required to *match* a designated position on the scale with his pointer, rather than to set it to a specified numerical value This limited phase of the problem was chosen because it permits data to be gathered rapidly and allows the accuracy of the setting to be objectively checked

The primary criterion employed is the *time* consumed in making a setting, since time is comparable from subject to subject, and from condition to condition In many of the experiments, action potentials from the active forearm were also picked up and measured However, action potentials cannot be compared from subject to subject, since it is not known that the efficiency of the pick up is the same in all subjects For any given subject they do provide at least a rough indication of the relative amount of muscular work involved under different conditions

APPARATUS

Figure 37 1 is an operational diagram of the essential mechanical features of the apparatus Rotation of the control knob turns the lower set of cone pulleys which

* Reprinted from *Journal of Applied Psychology*, Vol 33, No 4, August 1949

drives the upper set of cone pulleys through a belt Different ratios are ob tained by shifting the belt When the clutch is engaged, rotation of the upper shaft turns the drum and thus moves the pointer When the clutch is disengaged, movement of the knob does not affect the pointer

The linear scale consists of a black bakelite bar with vertical inserts of lucite 032′ wide at distances of 3, 9, 15, 21, 27, 33, 40, 56, 72, and 88 sixteenths of an inch symmetrically from the center Behind each insert is a tiny flashlight bulb

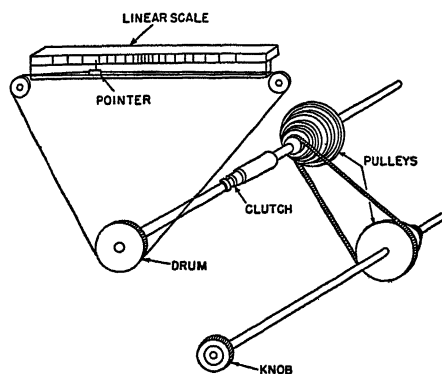



FIGURE 37 1 *Mechanical features—opera tional diagram*

Through the center of the linear scale runs a thin metal strip which is used in checking the accuracy of setting The pointer can be tipped to come in contact with the scale If the pointer is entirely within the limits of a lucite insert, it will not touch the metal strip If it is off the insert either way, it will come in contact with the metal strip and cause a red pilot lamp to light The limit of error tolerance is thus established by the width of the pointer

The mechanical system is without back lash and is so adjusted that the pointer remains exactly where it was set after the clutch is released To maintain these con ditions, the belt must be quite tight, so that there is noticeable resistance at ex tremely coarse ratios With the mechanical advantage of ratios in the medium and finer ranges, however, the operation re quires very little effort

For measuring time, two chronoscopes are used, so that time for travel to ap- proximate location and time for making the final adjustment can be separately de termined Similarly, two condensers are used to accumulate amplified action po tentials during the travel and adjust phases (Details of the electrical circuits and the four stage amplifier will be found in the Technical Summary Report of the project)[1]

PROCEDURE

The procedure was essentially the same for all experiments During a typical 2-hour experimental session 6 or 7 runs can be completed Each run consists of a series of 20 settings, involving all 20 of the lucite inserts in a scrambled order The proced- ure for a single setting is as follows

1 After giving a preliminary warning signal, the experimenter closes a switch which simultaneously (a) lights a pre selected lucite insert, (b) starts both chron- oscopes, (c) begins the accumulation of amplified action potentials in the first condenser

2 As soon as he sees an insert light up, the subject starts turning the knob to bring the pointer from the center of the scale to the designated position When the pointer comes within one tenth of an inch of the lighted insert, a contact is auto- matically closed which simultaneously (a) stops one chronoscope, (b) switches the accumulation of action potentials from the first to the second condenser Thus the first chronoscope and the first condenser provide measurement of the *travel* time and potential

3 When the subject has completed his setting, he pushes the clutch release with his non operating hand This action simul taneously (a) stops the second chrono scope, (b) cuts the second condenser out of the circuit Thus the second chronoscope

[1] W L Jenkins and M B Connor Op timal Factors for Making a Setting on a Linear Scale, Technical Report No 3, Con tract W28 099 ac 130, Watson Laboratories, Air Material Command, USAF, 30 June 1948 Restricted

and the second condenser provide the *adjust* measurements

4 The experimenter checks the accuracy of the subject s setting by tilting the pointer against the scale (Errors occur so rarely that the very occasional red light reading is simply discarded) The experimenter records the readings of both chronoscopes, and discharges each condenser separately into a sensitive ballistic galvanometer The apparatus can then be reset for another trial

Method of Analyzing Data

The raw data are in the form of time readings in tenth seconds and action potential readings in arbitrary meter scale units, for the travel and for the adjust phases of each setting The adjust readings cause no difficulty because they can be averaged directly However, travel readings vary according to the distance of the insert from the center Hence, travel readings are first plotted against distance traveled and a straight line fitted (The slope of this line is actually the travel rate, and the y intercept an estimate of the starting time or potential) Then the mean travel time (or potential) is scaled off for two standard distances 10 sixteenths and 50 sixteenths of an inch (The former is probably more representative of the usual amount of movement required in making discrete adjustments) Mean total time

TABLE 37 1

Influence of Ratio on Time and Potential

Standard Conditions

Mean Total Time

	10 Sixteenths Travel				50 Sixteenths Travel			
Ratio	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
220	25 2*	29 0*	24 0*	—	75 6*	66 6*	53 6*	—
454	17 5	24 1*	23 1*	35 1*	39 5*	42 9*	37 9*	72 3*
766	18 0	22 6*	22 4*	32 2	31 2*	35 4*	32 4*	52 7*
1 18	*16 3*	*19 5*	*19 4*	*30 3*	*24 3*	*22 7*	*25 8*	*44 3*
2 42	19 1*	21 6*	22 0*	29 1	27 1*	26 0*	25 6	38 7
4 08	19 2*	20 2	23 9*	35 4	23 6	24 6	27 9	42 2
6 28	19 5*	23 1*	26 7*	37 3*	23 5	27 5*	30 7*	43 3
9 70	23 8*	25 3*	28 1*	37 3*	26 6	28 9*	32 5*	42 1
16 3	32 8*	33 3*	37 2*	47 4*	34 4*	36 5*	42 4*	52 2*
33 6	54 3*	—	65 8*	—	57 9	—	73 0	—

Mean Total Potential

	10 Sixteenths Travel				50 Sixteenths Travel			
Ratio	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
220	24 3*	29 9*	26 9*	—	71 1*	78 7*	57 3*	—
454	16 8*	20 8	19 5	27 3*	41 6*	46 8*	36 7*	64 5*
766	15 3	19 5	19 0	22 1	28 5*	35 1*	29 4	42 9*
1 18	*14 4*	*19 7*	*20 3*	*20 3*	*23 2*	*28 1*	*28 3*	*36 8*
2 42	17 1*	16 4	21 2	17 5	25 1	22 6	26 8	26 3
4 08	16 5*	18 4	20 5	20 5	21 3	20 8	24 9	26 6
6 28	18 1*	16 4	21 9	25 8*	22 1	21 8	27 5	30 6
9 70	19 7*	18 4	22 6*	26 6	23 3	22 0	27 0	30 2
16 3	24 9*	23 4*	29 5*	33 4	26 9*	25 0	34 3*	37 8
33 6	25 4*	—	38 3*	—	28 1	—	43 9	—

* Significantly different from ratio 1 18

(or potential) = mean travel + mean adjust

Subjects

Two former Navy radar operators (DMS and HWQ) were used in all of the experiments Two other young men (JDS and RFM) with no such prior experience were available only for certain parts of the study These 4 subjects are right handed The young woman (JKD) used in the study is naturally left handed but was re quired to make settings with her right hand She also had had no particular me chanical background

Standard Conditions

The following conditions were standard in all experiments, unless specific excep tion is noted

Linear scale—At eye level and normal reading distance

Control knob—At waist level of seated subject, right hand operation, 2¾" di ameter knob

Error tolerance— 007" (pointer width of 025)

Ratios—Expressed in inches of pointer movement for one complete turn of the knob

Mean total time is expressed in tenth seconds for 10 sixteenths or 50 sixteenths travel distance Mean total potential is ex pressed in meter scale readings which have no absolute significance but are comparable for different conditions in the same sub ject Each mean is based on a minimum of 80 readings In tables showing italicized values an asterisk (*) indicates figures which differ significantly from the itali cized values, beyond the 1 per cent level of confidence

Results

Influence of ratio Is there an optimal ratio? Table 37 1 shows mean total time and mean total potential for ten ratios vary ing from 220 to 33 6 inches of pointer movement for one complete turn of the control knob Although the subjects differ in their general levels it is evident that the optimum is in the neighborhood of 1 18 in terms of both time and potential

Figure 37 2 shows why the optimal ratio is in this region For all subjects, travel time declines rapidly with increasing coarseness to about 1 18, thereafter coarser ratios do not speed up travel materially In the opposite fashion adjusting time de-

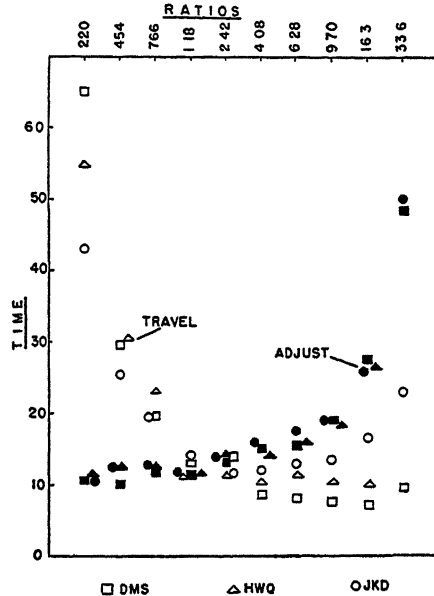



FIGURE 37 2 *Influence of ratio—standard conditions*

clines with *decreasing* coarseness of ratio to about 1 18, thereafter finer ratios do not aid in making the final adjustment A ratio about 1 18 combines rapidity of travel with speed of final adjustment

For convenience in the remainder of this report we shall refer to 1 18 as 'the optimal ratio' This should not be inter- preted too literally Actually there is an optimal *region* which holds good for all the subjects tested Well practiced subjects can use coarser ratios without undue loss, but the ratio designated as optimal has proved satisfactory for novice and expert alike

An indication of the stability of the op- timal ratio over a period of time is pre- sented in Table 37 2, which shows data for

TABLE 37 2

Stabılıty of the Optımal Ratıo

			Mean Total Tıme 10 Sıxteenths Travel								
	Subject DMS						*Subject HWQ*				
Ratıo	Feb 47	Apr 47	May 47	Oct 47	Mar '48	Ratıo	Feb '47	Apr 47	May 47	Oct 47	Mar 48
220	28 1	—	20 6	25 2	—	220	29 0	—	33 9	29 0	—
454	20 3	—	22 4	17 5	—	454	22 4	—	25 8	24 1	—
766	18 3	—	20 7	18 0	23 2	766	20 4	—	25 7	22 6	27 6
1 18	16 9	20 2	18 4	16 3	20 5	1 18	20 3	24 3	23 7	19 5	22 3
2 42	16 7	24 1	22 1	19 1	20 3	2 42	20 3	28 9	23 1	21 6	21 1
4 08	18 4	26 5	22 7	19 2	22 6	4 08	20 7	26 2	25 3	20 2	23 5
6 28	19 8	27 7	21 7	19 5	24 8	6 28	24 4	33 3	28 5	23 1	25 5
9 70	21 8	29 0	—	23 8	27 2	9 70	25 7	34 6	—	25 3	31 4
16 3	28 9	—	—	32 8	—	16 3	30 8	—	—	33 3	—
33 6	51 5	—	—	54 3	—	33 6	50 6	—	—	—	—

Standard condıtıons except that Feb '47 figures were obtaıned wıth a 2′ dıameter knob

two subjects gathered on five dıfferent oc casıons over a perıod of thırteen months Although the level of performance fluc tuates from tıme to tıme, the optımal ratıo remaıns ın the same regıon

Table 37 3 shows that the optımal ratıo holds good for both the domınant and non domınant hand (To obtaın these figures, a left hand and a rıght hand knob were coupled wıth auxılıary belts, so that the poınter could be set wıth eıther hand) Partıcularly ınterestıng here are the data for subject JKD Although naturally left-handed, JKD had by thıs tıme become well practıced ın rıght handed operatıon of the apparatus At unfavorably hıgh ratıos she was now able to make faster settıngs wıth her rıght hand Around the optımal ratıo, the two hands were equally good

Influence of knob dıameter In a pre lımınary study on two subjects, 14 knob dıameters were tested wıth 5 dıfferent ratıos For clarıty ın presentatıon the 14 dıameters are grouped ın five step ınter vals Table 37 4 gıves the mean total tıme

TABLE 37 3

Ratıos ın Rıght vs Left Hand Operatıon

			Mean Total Tıme 10 Sıxteenths Travel			
	DMS		*HWQ*		*JKD*	
Ratıo	Rıght	Left	Rıght	Left	Rıght	Left
766	22 2	24 4	25 5	29 5	25 0*	24 9
1 18	*21 3*	*24 6*	*24 8*	*28 0*	*22 5*	*23 6*
2 42	21 0	24 3	24 7	26 1	24 7	22 4
4 08	23 7*	25 0	25 1	26 4	27 6*	30 3*
6 28	26 0*	29 3*	29 8*	31 5*	27 7*	33 7*
9 70	29 2*	38 0*	37 6*	38 6*	31 6*	36 4*

Standard condıtıons except that ıdentıcal rıght and left hand knobs were coupled by a belt so that eıther could be used

* Sıgnıficantly dıfferent from ratıo 1 18

for 10 sixteenths travel distance Several points of interest appear (1) Regardless of knob diameter, the optimal ratio remains in the neighborhood of 1 18 (2) It is apparently not possible to compensate for an unfavorable ratio by altering the size of the control knob Notice that the fastest times for ratio 6 28 are longer than the slowest times for ratio 1 18 (3) With coarse ratios the larger knob diameters work better (4) At the optimal ratio, knob diameter appears to make very little difference

As a check on this last point, 5 knob diameters were studied at the optimal ratio, using 4 subjects Table 37 5 shows the results for both time and potential In terms of mean total time, only the half inch diameter is clearly unfavorable for all subjects, and the one inch diameter mildly so for two of them In terms of action potential, the 2¾" diameter is significantly superior to the smaller sizes, although not always to the 4" diameter

Figure 37 3 shows travel time and adjusting time separately The half inch diameter yields longer times for both travel and adjusting in all subjects Among the

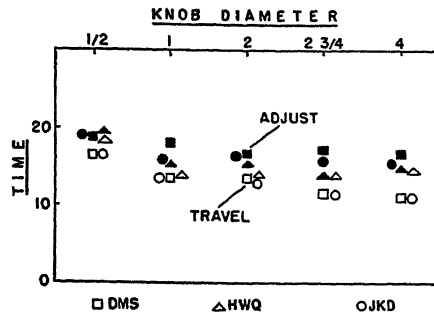



FIGURE 37 3 *Influence of knob diameter—standard conditions*

larger sizes there is little to choose It appears that the critical motion is the twist of the forearm, not the movement of the finger tips Practically speaking, as long as the optimal ratio is used, the exact knob diameter does not matter, unless it is too small or too large to be grasped conveniently The standard 2¾" size used in most of our experiments was adopted simply because most subjects expressed a preference for this size

TABLE 37 4

Interaction of Knob Diameter and Ratio

Mean Total Time
10 Sixteenths Travel

Subject HWQ

Knob Diameters	Ratio 1 18	Ratio 2 42	Ratio 4 08	Ratio 6 28	Ratio 9 70
½, ¾	29 2	—	—	46 4	—
1, 1¼, 1½	24 1	26 8	26 8	35 1	40 2
1¾, 2 2¼	22 6	25 3	25 6	31 2	34 2
2½, 2¾, 3	23 6	27 0	25 7	31 6	33 0
3¼, 3½, 4	24 3	27 3	25 0	30 8	30 7

Subject DMS

Knob Diameters	Ratio 1 18	Ratio 2 42	Ratio 4 08	Ratio 6 28	Ratio 9 70
½ ¾	21 5	—	—	34 1	—
1, 1¼ 1½	21 5	24 3	30 0	37 5	33 7
1¾, 2 2¼	22 5	22 2	26 6	34 5	28 3
2½ 2¾, 3	21 6	22 5	26 3	28 4	29 6
3¼, 3½, 4	23 2	22 4	25 7	29 9	27 2

Standard conditions except that series of knob diameters were combined with series of ratios as indicated

TABLE 37 5

Influence of Knob Diameter at Optimal Ratio

	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
Diam	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
½	25 3*	28 1*	26 3*	42 1*	35 3*	38 1*	35 5*	53 7*
1	23 1	23 0*	22 0	39 3*	31 5	29 4	29 6	46 1*
2	21 1	22 9*	23 0	35 2	30 3	29 3	29 4	44 0*
2¾	*21 9*	*20 8*	*22 1*	*34 5*	*28 7*	*28 0*	*27 3*	*38 9*
4	21 2	21 8	21 8	37 6	27 6	29 4	26 6	46 6*

	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
Diam	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
½	31 4*	29 2*	38 0*	33 4*	44 6*	40 0*	50 8*	46 2*
1	30 9*	24 2	33 0*	27 1*	44 1*	35 0	44 2*	35 9*
3	26 0*	25 5*	27 6	22 3*	38 4*	36 3*	39 2*	32 3*
2¾	*23 4*	*22 6*	*26 0*	*18 5*	*33 0*	*33 8*	*36 8*	*22 5*
4	21 7	26 7*	24 9	13 6	31 7	37 5	35 2	19 6

Standard conditions except that series of knob diameters were combined with ratio of 1 18

* Significantly different from 2¾

Influence of crank handle Cranks are generally used in tracking operations The question has been raised whether a crank is better than a knob for making discrete settings involving large amounts of travel To study this problem the 2¾ knob was drilled so that a crank handle could be attached ¼′ from the periphery Time measurements were taken at 7 ratios under the following conditions (1) Knob alone as a control, (2) crank attached and its use required, (3) crank attached but its use optional

Table 37 6 shows mean total time for 50

TABLE 37 6

Comparison of Knob and Crank

	Mean Total Time								
	50 Sixteenths Travel								
	Subject DMS			Subject HWQ			Subject JDS		
Ratio	KNOB	CRANK	OPT	KNOB	CRANK	OPT	KNOB	CRANK	OPT
220	81 2	52 6	54 8	73 5	58 5	55 1	103 6	50 1	52 8
454	52 7	35 6	36 7	48 0	42 5	39 9	64 6	40 1	38 7
766	37 7	30 2	31 0	40 3	39 9	35 4	45 3	33 4	29 0
1 18	25 6	32 7	32 5	30 6	38 6	31 7	29 0	34 3	32 7
2 42	26 0	33 5	26 7	27 8	39 8	36 2	29 8	36 2	32 1
4 08	26 8	45 8	29 6	30 0	45 6	34 0	29 0	44 0	32 0
6 28	24 6	43 8	29 1	32 8	61 7	32 7	31 2	43 7	33 1

Standard conditions except each mean based on a minimum of 40 readings Crank simu lated by attaching crank handle to periphery of 2¾″ knob In the table KNOB means knob alone CRANK means use of crank required OPT means crank handle present but use optional

sixteenths travel distance, which should give the crank the maximum advantage Two interesting points appear (1) Although the crank speeds up setting at ratios below 1 18, it does not enable these ratios to compete with the optimal ratio and the simple knob (2) At the optimal ratio, the forced use of the crank is definitely deleterious and even its mere presence appears to hamper the best performance Within the limitations of these experiments at any rate, it appears that a crank handle serves no function whatever in making discrete settings on a linear scale

Influence of backlash Backlash is unavoidably present in some equipment What is its influence on the speed of making settings? To study this question, the apparatus was modified by the addition of an arm moving between adjustable stops immediately beyond the subject s control knob, so that varying degrees of backlash could be introduced In a preliminary series with two subjects backlash was tested in 1 steps from 0° to 20 in the expectation that some particular amount of backlash might prove to be critical Since this expectation was not realized, the figures have been grouped into 7 step intervals Table 37 7 shows mean total time for 10 sixteenths travel at ratios 1 18 and 6 28 Surprisingly, backlash appears to have very little effect, even at the unfavorably coarse ratio of 6 28

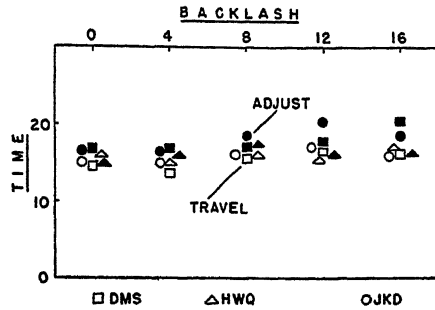



FIGURE 37 4 *Influence of backlash—standard conditions*

As a further check, backlash of 0 , 4 , 8 , 12 , and 16 was tested with 3 subjects using the optimal ratio Results are given for mean total time and mean total potential in Table 37 8 Again it seems that no substantial effect of backlash can be found in either time or action potential There is a slight upward trend with increasing backlash, but the statistically significant differences are scattered spottily and unconvincingly throughout the table Figure 37 4 indicates that the slight upward trend comes from a minor lengthening of adjusting time, while travel time remains unaffected

We are reluctant to draw the sweeping conclusion that backlash is totally unimportant under all conditions Perhaps with

TABLE 37 7

Interaction of Backlash and Ratio

	Mean Total Time 10 Sixteenths Travel Subject DMS		Subject HWQ	
Backlash in Degrees	Ratio 1 18	Ratio 6 28	Ratio 1 18	Ratio 6 28
0, 1 2	23 1	27 8	24 4	29 2
3, 4, 5	23 2	30 1	24 9	28 1
6 7, 8	23 8	32 5	25 8	28 7
9, 10, 11	25 4	33 0	26 4	30 1
12, 13 14	25 1	32 7	26 4	32 2
15, 16, 17	26 1	32 5	26 2	30 7
18 19, 20	26 5	33 3	26 6	29 7

Standard conditions Varying degrees of backlash introduced by means of an arm working between adjustable stops, immediately beyond subject s control knob

TABLE 37.8

Influence of Backlash at Optimal Ratio

Back-lash	*Mean Total Time* 10 Sixteenths Travel			50 Sixteenths Travel		
	DMS	HWQ	JKD	DMS	HWQ	JKD
None	*21.9*	*22.9*	*23.7*	*31.1*	*30.9*	*31.7*
4°	22.0	23.8	23.4	30.4	31.0	31.4
8°	23.4	25.5*	26.6	32.6	33.5	34.6
12°	24.2*	24.1	28.6*	34.2*	31.7	37.4*
16°	26.8	24.5	26.6	36.4*	33.3	34.6

Back-lash	*Mean Total Potential* 10 Sixteenths Travel			50 Sixteenths Travel		
	DMS	HWQ	JKD	DMS	HWQ	JKD
None	*25.7*	*23.9*	*32.9*	*38.9*	*36.7*	*43.7*
4°	28.0*	24.2	31.2	40.8	36.2	42.0
8°	26.4	26.6*	32.7	39.2	37.0	45.5
12°	26.7	25.3	33.9	39.5	37.3	46.3
16°	29.0*	26.5*	32.7	42.2	37.7	45.5

Standard conditions. Varying degrees of backlash introduced by means of an arm working between adjustable stops immediately beyond subject's control knob.

* Significantly different from None.

excessive friction or inertia, perhaps when far greater accuracy than .007″ is demanded, backlash may prove more disturbing than in the present experiments. Those are questions for further research to answer.

Influence of error-tolerance. How much does it slow up an operator to demand greater accuracy in setting? In our apparatus the error-tolerance could be altered simply by changing the width of the pointer in relation to the width of the lucite inserts. In a preliminary series, 11 pointer-widths were tested. Table 37.9 shows the results in terms of mean total time for 10 sixteenths travel distance. At the optimal ratio, only subject DMS shows a marked lengthening of time with decreas-

TABLE 37.9

Interaction of Tolerance and Ratio

Error Tolerance	*Mean Total Time* *10 Sixteenths Travel* Subject: DMS		Subject: HWQ		Subject: JDS	
	Ratio 1.18	Ratio 6.28	Ratio 1.18	Ratio 6.28	Ratio 1.18	Ratio 6.28
.018″, .016″	17.0	17.8	19.5	21.7	—	—
.013″, .011″	16.5	21.0	20.7	23.2	22.1	27.9
.009″, .008″	18.2	24.4	22.7	27.7	22.7	30.1
.007″ .006″	24.2	28.6	22.7	30.0	24.2	30.0
.005″	26.5	37.1	24.1	29.5	29.9	33.4
.004″	30.0	52.1	25.2	33.2	32.7	39.9
.003″	35.3	50.2	29.0	39.2	33.9	40.5

Standard conditions except that knob diameter is 2″.

TABLE 37 10

Influence of Tolerance at Optimal Ratio

	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
Toler	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
012′	15 8*	19 0*	16 6*	27 9*	22 8	25 4*	23 0	38 3*
009	17 1	19 5*	18 3	31 4	23 9	26 3	24 7	40 2
007	*17 5*	*22 6*	*19 8*	*34 6*	*24 7*	*27 8*	*25 0*	*45 0*
005	20 7*	23 4	21 8*	38 1	27 9	31 0*	27 4	48 9
003	27 7*	30 4*	25 9*	51 6*	33 3*	37 2*	32 3*	61 6*

	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
Toler	DMS	HWQ	JKD	RFM	DMS	HWQ	JKD	RFM
012″	14 1*	14 3*	21 4*	19 6	22 1*	22 7	30 6	30 0
009	14 9	15 7	22 0*	19 5	23 2	22 9	31 6	29 9
007	*15 9*	*15 8*	*24 8*	*21 6*	*24 3*	*23 0*	*33 6*	*32 8*
005	17 2	19 6*	23 4	23 4	25 2	27 2*	31 8	34 6
003	21 5*	23 7*	27 6*	27 2*	29 0*	30 5*	36 4	37 6*

Standard conditions except that series of error tolerances were tested at ratio of 1 18
* Significantly different from 007

ing tolerance, but at ratio 6 28 all 3 subjects show the same effect

A further study was made with 4 subjects, using 5 tolerances at the optimal ratio, measuring both time and potential Table 37 10 gives the results There is evidence of a moderate lengthening of time from 012″ to 005′, then a sharp break at 003″ From the reports of the subjects, it appears that 003′ represents a breaking point at which it becomes perceptually impossible to judge whether the pointer is accurately positioned This is borne out by the fact that only at this level of tolerance did the subjects have an appreciable number of red lights (indicating that the clutch was released when the pointer was not within the confines of the lucite insert)

Figure 37 5 shows, as might be expected, that error tolerance does not affect travel time Adjusting time increases slowly as tolerance decreases, with a sharp upward break at 003′

It should be realized that 003″ represents a perceptual limit only under the conditions of this experiment, i e, centering a pointer of appreciable thickness on a lighted insert With ideal conditions, such as a fine hair line, it might be expected that the perceptual limit would be considerably lower

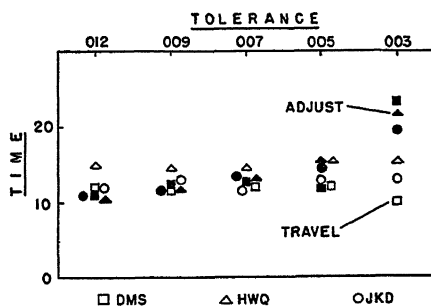



FIGURE 37 5 *Influence of tolerance—standard conditions*

SUMMARY

In the foregoing experiments, the subject was required to move a pointer by means of a control knob and set it to a position on a linear scale indicated by a lighted insert Time consumed in making the setting and the relative action potential

developed in the active forearm were measured separately for travel to approximate location and for final adjustment Systematic variations in ratio, knob diameter, backlash, etc, were introduced Three to 5 subjects were used in the various parts of the study The principal results follow

1 The optimal ratio is 1 or 2 inches of pointer movement for one complete turn of the knob, for either the dominant or non dominant hand Finer ratios waste time and effort in traveling to the approximate location Coarser ratios are clumsy for making the final adjustment No other design factor investigated is as important as the optimal ratio

2 Knob diameter is relatively unimportant, as long as the knob is large enough to be grasped conveniently An unfavorably coarse ratio cannot be compensated for by altering the size of the control knob

3 An unfavorably fine ratio cannot be compensated for by substituting a crank handle for the control knob When the optimal ratio is employed, the addition of a crank handle to the knob does not aid and may be actually harmful, even when its use is optional

4 Backlash even in excessive amounts, has a relatively minor influence on either time or potential at the optimal ratio—under the conditions of this experiment This may not be true under conditions of extreme friction and inertia, or when a tolerance much finer than 007" is required

5 Demanding greater accuracy of the subject by reducing the permitted error-tolerance increases time and potential only moderately, as long as the optimal ratio is employed The final limit of accuracy in the present experiments appeared to be set by the perceptual difficulty of centering a pointer of appreciable thickness on a lighted insert, rather than by the limits of motor control

Psychological Aspects of Stick and Rudder Controls in Air Craft *

JESSE ORLANSKY

This paper is an abridged version of Report No 151 1 8 submitted to the Special Devices Center, Office of Naval Research under contract N6ori 151 with The Psychological Corporation

SUMMARY

This paper is an attempt to determine how airplane control systems may be designed to provide the pilot with optimal sensory information by means of pressure cues obtained from operating the stick and rudder The present approach to the problem consists of an examination and evaluation of literature pertaining to

(a) The maximum forces that may be exerted by a human pilot

(b) Human reaction time insofar as it may be expected to cause delays in the pilot s response

(c) The optimal design, placement, and manner of movement of controls

(d) The optimal gradient of control forces

Certain recommendations for aircraft control systems are discussed

STATEMENT OF PROBLEM

In this paper, an attempt is made to examine some of the psychological problems associated with the operation of an airplane It is a recognized fact that the

* Reprinted from *Aeronautical Engineering Review* Vol 8, No 1, January 1949

control of an airplane may impose requirements beyond those that can be met by a human pilot Even in routine operations, it is desirable that the arrangement of the cockpit, the controls, and the instruments (as well as their influence upon coordination between crew members) be such as to provide an accurate and easy flow of pertinent information This is construed to be the psychological consequence of the mechanical design embodied by a particular airplane Attention is directed, in the present study, to some of the human factors associated with the stick and rudder controls as they might be found in a high-performance fighter type airplane Though not attempted here, similar treatment might well be accorded to the instrument panel, to the radio console, to the flight engineer's desk, and, especially, to the integration of duties by the crew members

The handling of high speed aircraft requires the control of enormous forces by the application of equal counterforces, only a part of which can be supplied by the pilot Since aerodynamic pressures increase markedly with speed, while the pilot's strength is relatively fixed, some means must be employed to assist the pilot in moving the control surfaces on the newer airplanes

Conventional control linkages permit the pilot to perceive some of the airplane's flight characteristics through position and pressure effects on the stick and rudder controls These effects are called "stick (or rudder) feel," and pilots rely upon them in flying the airplane "Stick feel" depends, in part, on the cues arising from the feedback of some fraction of the aerodynamic forces developed upon the control surfaces Mechanical boosters introduce special "feels" on the controls caused by friction, time lag, pulsation, inertia, and other attributes of the booster system Thus, as the fraction of force supplied by the pilot diminishes, feel becomes more and more dependent on the operating equipment rather than on flight conditions Some modern planes employ mechanisms with a booster to pilot force ratio of 10 1, (i e, the pilot supplies only one tenth of the required control force), while future designs may require ratios as high as 1,000 1

At the present time, pilots have come to expect certain stick feel effects as the control stick is moved to various positions at various speeds Booster mechanisms may so modify this relationship that stick feel varies almost independently of control surface pressures In one system, for example, displacement of the control stick is related directly to displacement of the control surface, stick pressure remains constant at a low value, thereby eliminating any possible differential pressure cues Or, the stick may be used to initiate control surface motion, while the degree of stick deflection monitors the rate of change In this case, again, the normal pressure and displacement cues are altered "Normal" control feel may be re established by artificial means but, to accomplish this it would first be necessary to understand what is meant by "normal" control feel Current specifications do not rigidly determine a standard control feel, and it can be shown that current airplanes actually differ in their feel characteristics

An airplane may be flown without "normal" control feel, as evidenced by the operation of remotely controlled aircraft However, it is not yet feasible to maneuver in this manner a fighter aircraft, as if in combat Jet fighter pilots, 15 of whom were interviewed in connection with this project, indicate that there is time for only slight attention to the instruments during high speed acrobatics in planes like the F 80 Shooting Star and F 84 Thunderjet During maneuvers, these pilots maintain their primary orientation by reference to the horizon and stick feel, with secondary regard to three of the instruments the Machmeter, yaw indicator, and altimeter They regard stick feel as a particularly valuable cue because it is always available without distracting the pilot's attention from his target A pilot upon whom are placed the tasks of navigation, communication and aerology in addition to flight control and combat, approaches the limit of his abilities For such a man, a stick with feel is equivalent to a host of flight instruments

Since current airplanes are not consistent in their control feel characteristics, present practice alone does not dictate a desirable

standard for future aircraft It would appear useful to examine several questions generally applicable to all airplanes regardless of their speed

(a) What are the maximum forces that may be exerted by a human pilot?

(b) What delays may be expected as a consequence of the pilot s reaction time?

(c) Where should the controls be placed, and how should they move for most efficient manipulation by the pilot?

(d) What gradient of stick forces will provide the pilot with optimum pressure cues?

EVALUATION OF HUMAN CAPACITIES FOR AIRCRAFT CONTROL

(A) The maximum forces that may be exerted by the pilot It is obvious that the maximum control forces required of a pilot must never exceed the limit of his strength These forces are limited to the following maxima on stick type controls

elevator	35 lbs
aileron	30 lbs
rudder	180 lbs

These values appear to be based on a study of two pilots carried out in 1936 by

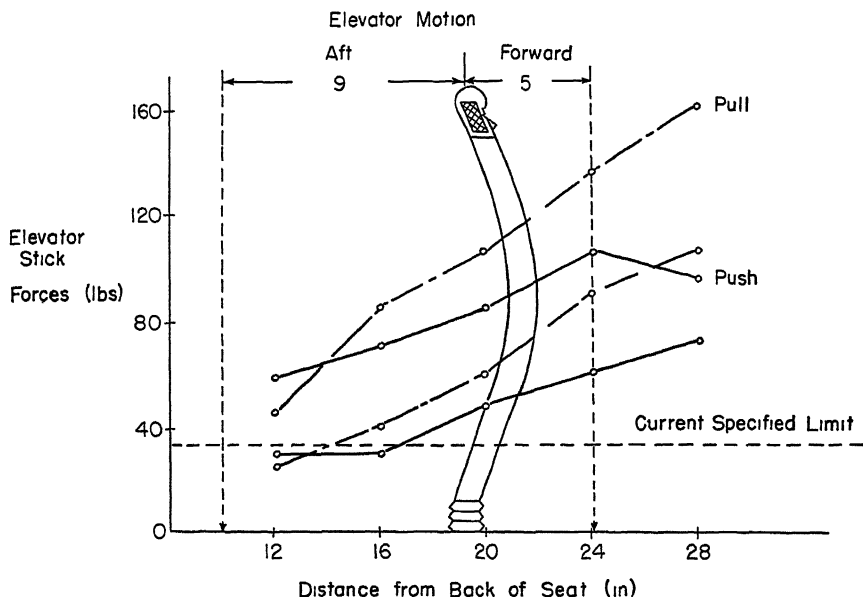



FIGURE 38 1 *Maximum push and pull forces exerted on elevator at various hand positions (lesser force of two pilots) The two upper curves represent performance in the most favorable while the two lower curves represent the least favorable lateral position (range ± 8 in from center) The limits of elevator motion in a standard cockpit and the maximum permissible elevator force fixed by specification (35 lbs) are indicated*

In this paper, the problem is approached by an examination of published information and by extensive interviews with jet plane pilots The study indicates a direction for the experimental work which may be desirable to verify the present conclusions Attention in this study is directed primarily to fighter aircraft equipped with conventional stick and rudder controls

the National Advisory Committee for Aeronautics (9) A comparable study for wheel type controls was reported in 1937 (19)

(1) Elevator control force Fig 38 1 is derived from data presented in the N A C A report The two upper curves represent the maximum push and pull forces that may be exerted in the most favorable

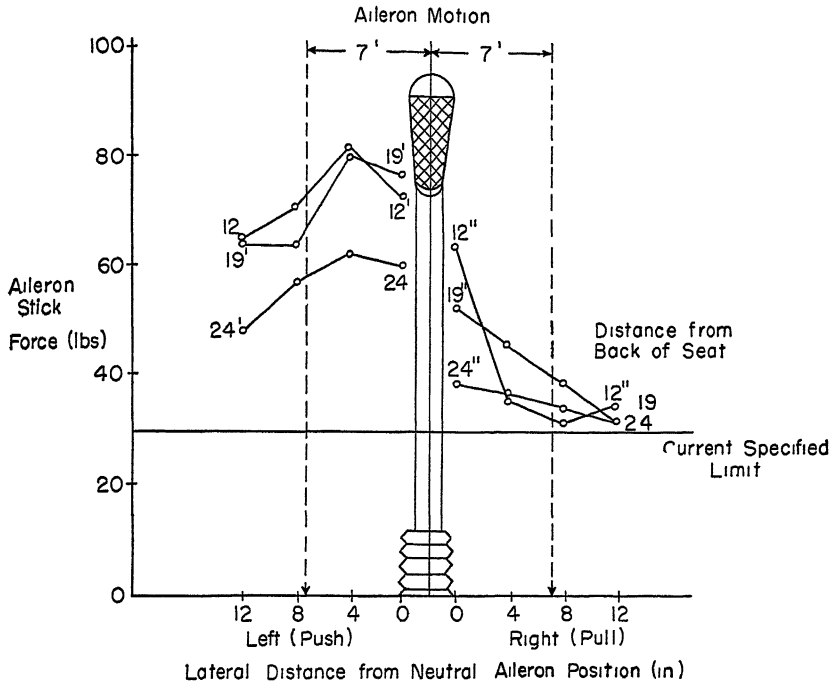



FIGURE 38 2 *Maximum aileron forces exerted at various hand positions by one pilot The limits of aileron motion in a standard cockpit and the maximum permissible aileron force fixed by specification (30 lbs) are indicated*

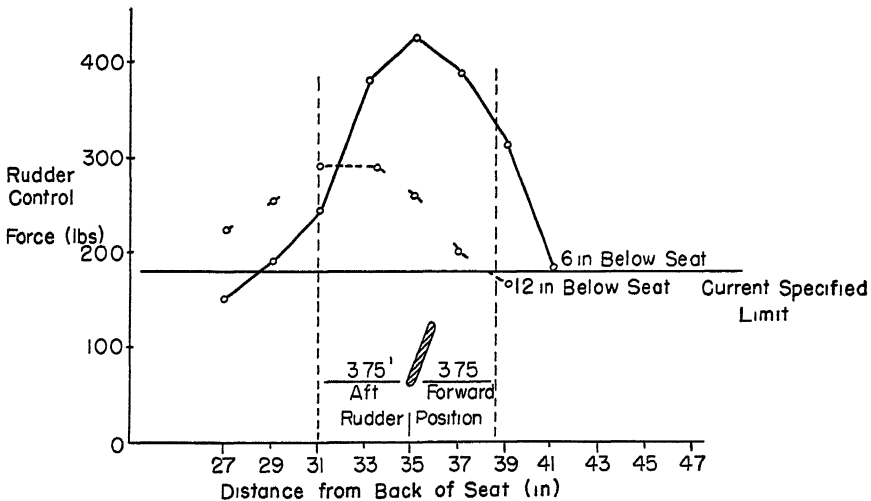



FIGURE 38 3 *Maximum rudder force (lesser force of two pilots) exerted at various foot positions with the rudder bar at two levels below the seat The limits of rudder motion in a standard cockpit and the maximum permissible rudder force fixed by specification (180 lbs) are indicated*

lateral position, which was right of neutral for these right handed pilots, the two lower curves represent the maximum forces in the least favorable lateral position It is clear that greater pull forces than push forces may be exerted in all hand positions except those close to the seat The ability to exert a push or a pull force increases with distance from the seat Accepting the two pilots as a representative sample, it would appear that pilots would always be able to exert the maximum allowable elevator type force except where the hand is close to the body

(2) Aileron Control Force The data for aileron forces are shown in Fig 38 2 The right handed pilot can exert greater aileron force to the left (i e, push) than to the right (pull) of neutral the ability decreasing with lateral and forward displacement The maximum aileron forces that can be exerted are less than the maximum elevator forces and do not vary so much with changes in hand position The data indicate the influence of hand position on the ability to exert aileron forces and show that performance decreases at extreme positions Apart from aerodynamic considerations, they also suggest that right-handed pilots might find it easier to perform counterclockwise rolls and turns to the left than clockwise rolls or right turns

(3) Rudder Control Force The N A C A data on rudder forces indicate that the design limit of 180 lbs can generally be exceeded, as based upon the maximum rudder forces exerted by the weaker of the two pilots Fig 38 3 also shows that rudder forces fall off sharply as the seat height increases above the rudder

Taken together, these 3 graphs suggest the following conclusions

(a) The force limits imposed by current design specifications are generally lower than the maximum forces that humans can exert

(b) The permissible aileron forces approach human limits for this type of motion

(c) There appears to be a reasonable margin between elevator and rudder forces and the human limits for these types of motion

It can hardly be doubted that the standard or maximum allowable forces are based on an inadequate sampling of the pilot population One consequence of this fact is that there is inadequate knowledge of the safety factor allowed by the present standards Pilot acceptance of present control force standards might be interpreted as a demonstration of their validity, however, pilots often have endured undesirable practices without objection Apart from the maximum allowable force, another requirement is that the actual force expenditure be optimum to minimize fatigue and to facilitate delicate control adjustments The next section is devoted to this consideration

(B) Sensory discrimination of control pressures Various control motions are required for take off, maneuvering, and landing, and, as has been shown, the required forces should not be excessive An important psychological question is whether these forces increase by magnitudes that permit the pilot to make his most sensitive adjustments A pilot cannot detect changes of a few ounces in the pressure—i e, feel" —of the controls, nor, while exerting a force of 100 lbs, could he detect an increase of 1 lb There is probably an optimum pattern of pressure increases which would furnish the pilot with a maximum number of discriminable cues

This consideration relates to the Weber Fechner law, a famous psychological generalization, first stated in 1834, on the perception of differences—i e, human sensitivity As Woodworth puts it, 'in comparing magnitudes, it is not the arithmetical difference but the ratio of the magnitudes, which we perceive" (28) The significance of this generalization, insofar as it applies here, is that one should not expect a pilot to detect the same differences in pressure at all points in the pressure continuum He might, for example, discern a difference between 5 and 6 lbs (ΔI* equals 1 lb), but require an increase from 15 to 18 lbs (ΔI equals 3 lbs) before he could again note a difference That is,

$$\Delta I / I = k$$

* ΔI represents the discernible increment in intensity or the just noticeable difference

where ΔI is the just discernible increase in intensity I and k is a constant In the examples just given, $\Delta I/I = 1/5$ and 3/15 and $k = 20$ per cent Intensity perception is relative and not absolute

An investigation of pressure discrimination has been carried out by Jenkins (15) at the Aero Medical Laboratory, Wright Field A cockpit mock up was prepared so that the accuracy of reproducing the various types of control pressures on stick, wheel, and rudder could be determined specified pressures and the averages actually attained were measured in pounds and called the constant errors The closer a value is to zero constant error, the more accurate is the performance The data show that pilots tend to overexert when trying to push (or pull) small pressures, while they underexert for the larger pressures One may observe in Fig 38 4 (based on Jenkins study) that more sensitive control is possible by means of the stick than by either wheel or rudder At all pressures

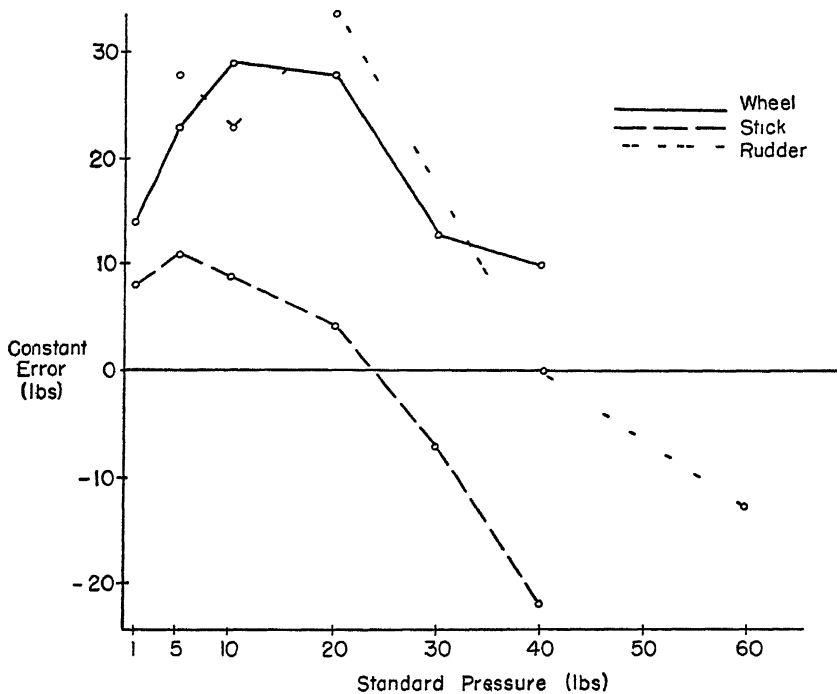



FIGURE 38 4 *Accuracy of performance in applying certain specified pressures to stick, wheel and rudder controls*

The subjects were blindfolded and, after practice, were required to apply designated pressures on the controls By this technique data were gathered on the accuracy and consistency of performance of 20 AAF pilots and 13 nonpilots No information was collected on discrimination of angular displacement of the stick or on a flight simulating task requiring continuous adjustment

The differences between the standard up to 30 lbs, the constant errors are least for the stick control, with wheel and rudder following in that order This is more sharply indicated in Fig 38 5, which shows relative accuracy as determined by the ratio of constant errors to the standard pressures The lower the ratio, the more accurate the performance The stick is, of course, the most accurate control agent among the three types considered, and its relative accuracy is fairly constant from 5

to 40 lbs, no difference was found be tween the accuracy of elevator and aileron type motion, the relative accuracies of the wheel and rudder are constant from about 15 to 60 lbs, the largest value tested

The following conclusions may be drawn from Jenkins study

(a) A pilot will be able to discriminate more pressure cues if stick pressure in creases in a nonlinear (rather than linear) manner with respect to its independent variable, such as stick displacement or air plane speed

FIGURE 38 5 *Relative accuracy of performance in applying pressures to stick, wheel, and rudder controls*

(b) Control pressures should occur over a wide range in order to provide the pilot with as many perceptible pressure differ ences as possible

(c) When control pressures are low, they provide poor cues They should rarely be less than 5 lbs This requirement would also appear to be necessary to overcome the masking effect due to friction Merely resting the hand on the stick results in some pressure because of the weight of the arm, the same is true for the rudder pedals, where the average pressure due to the weight of the feet was found to be 7 lbs

(d) When attempting to exert a small force, the individual tends to apply a greater force than is required Conversely, he underexerts when a large force is re quired There is, therefore, an optimum range for control forces which may be esti mated as 5 to 30 lbs for elevator and aileron and 7 to 60 lbs for rudder

(e) Pilots appear to be more accurate than nonpilots in these tests The number of flying hours and body weight were not related to accuracy Performances im proved with practice and with knowledge of results When a light force imme diately follows a heavy one (or vice versa), there is some evidence that the accuracy of a performance is adversely affected

Pilots' opinions concerning the stick forces they have exerted in flight show that they are apt to be inaccurate judges Thus, Gough and Beard (9) report that two ex perienced test pilots made estimates that were found to be in error by as much as 50 per cent when checked against instru ment records They were most accurate in reporting pressures of about 10 lbs, they exerted more force than they thought they did in the case of small values and less in the case of large values De Beeler (3)

showed that pilots vary considerably in reproducing in a mock up the rate of motion they would use to pull out of a dive The present author has examined records that show that pilots actually exerted only 40 to 50 lbs during flights on which they reported they had exerted 100 lbs

A number of investigators have examined the factors that influence accuracy in the operation of hand controls Their interest has generally been directed at manual controls for tanks and guns, but some of the findings are applicable to the present topic Craik and Vince (5–7) report that friction of approximately 2 lbs in a hand control is desirable to eliminate the effects of body sway, hand tremor, jolting, and vibration to protect the operator against involuntary sagging of the arm, as well as to smooth out control movements Performance was more accurate when visual observation of an instrument display was permitted, in addition to detection of the pressure cues

For precision of adjustment, Hick (11) advises no control motion below the limits of 2 lb pressure and 2 in movement His experiments, as do those of Craik and Vince, also show that small forces and distances are overestimated, while large forces and distances are underestimated Errors of 5 to 15 per cent are found in the manual exertion of force (13) A pressure gradient with velocity led to an improvement in handle winding performance According to Hick (12), friction (up to 4 lbs) at the handle reduces average error by about 15 per cent under conditions of jolting but is unfavorable when no jolting is present

One may conclude on the basis of these studies that

(a) The perception of changes in pressure, such as observed in airplane control systems, is not an absolute ability but is relative to the level of pressure at which the change occurs The increments of stick pressure in response to changes in stick displacement or speed should be geometric rather than arithmetic in order to furnish the pilot with the maximum number of discriminable pressure cues

(b) The pilot is most sensitive to pressure differences when controls are operated against a moderate work load The optimum range of this load for accuracy and consistency of performance is of the order of 5 to 30 lbs for stick and 15 to 60 lbs for wheel and rudder controls (higher values were not tested for the two latter controls) Higher loads would probably increase fatigue to an undesirable degree

(c) Some friction on the controls is advantageous in eliminating the effects of hand tremors, jolting, and vibration because it tends to smooth out motion The level of desirable friction on hand control is reported variously as 2 to 5 lbs While there are no data on desirable rudder pedal friction, there is a hint that it should be of the order of 7 lbs, as judged by the average pressure exerted by the resting weight of the foot

(C) The position of controls and the direction of motion The advent of power operated controls permits the design of controls in any size, shape, and position deemed desirable for ease of performance An evaluation of novel type controls may be recommended, but it is beyond the scope of this paper However, attention should be directed to the effect upon performance of such factors as direction of movement, size, shape, and position of the controls

Considerable anthropometric data are now available on the population likely to operate airplanes (8, 22, 23), tanks (1), and similar military equipment The dimensions of the standard cockpit are based on such information Recently, King, et al (18), measured the functional reach of 139 young males, of whom 79 were Navy pilots He found that the stick and rudder in the standard cockpit could be manipulated to their limits of motion by 97 per cent of that population, but 3 per cent would have difficulty

It may be expected that the precision of linear adjustment, such as required on stick and rudder controls, varies somewhat with the position of the hand and foot King remarks that the precision of movement of the hand and fingers decreases as an unsupported arm is extended' None of the available investigations, however, gives quantitative measures of the accuracy of manual (or pedal) control motion for various positions and distances similar to

what Jenkins has done for control pressures. Ideally such investigation would reveal the distance through which the hand (or foot) must move at various extensions and under various loads, before a just noticeable increment occurs.

Vince (25) shows that the direction of control motion should be similar to the expected direction of its effect, especially for performances requiring rapid adjustments. This finding, which is confirmed by Warrick (26), is of special applicability in airplanes, where rapid adjustments of controls are so frequent. With further development of high speed aircraft, the importance of relating direction of control motions to direction of effects will increase tremendously.

Grether (10) tested the relative efficiency of several types of aircraft control motion in a simple pursuit task. The subjects (24 nonpilots in one experiment, 36 rated pilots in three other experiments) were required to move each control so that a pointer randomly activated returned to its reference mark. The efficiency of performance was measured by a clock that cumulated the time intervals during which the pointer was kept within the reference mark. Five control motions—i.e., rudder, stick aileron, wheel aileron, stick elevator, and wheel elevator—were studied. Various comparisons were made of such conditions as equal or unequal extent of control motion and angle of knee or arm flexion on the controls. Grether concludes that

(a) Hand controls (stick or wheel) are better than foot controls (rudder), for equal and unequal extents of movement.

(b) Elevator movements (fore and aft) are slightly better than aileron movements (lateral or rotary) on stick and wheel controls.

(c) The wheel and stick controls yield approximately equal efficiency for aileron and elevator type motion.

(d) There are differences in comfort but not in efficiency on tests performed under average leg and arm angles of 105°, 120°, and 135.

One study (14) employed 18 pilots to investigate the effect of offsetting the stick and rudder controls from their normal central positions. There is a strong tendency to pull the controls back to a laterally symmetrical position, while fore and aft motion does not appear to be affected. Control motion is most accurate when the position of the hand is at normal elbow height, and hand tremor increases appreciably when the hand is held more than 8 in. above or below the level of the heart (5). When the operator can observe visually the effect of his manipulations his accuracy of control is greater than when he is dependent on kinesthetic cues alone (2).

The shape of a handle affects the ease of control of machines and tools, but no studies are reported on the preferred shape of an airplane control stick. A shift from a round knob to a pistol grip control improved by about 8 per cent the tracking, ranging, and triggering performance on the B 29 pedestal gun sight (16). The diameter of a hand grip should be approximately 1.5 in. and should provide friction (e.g., be rubber covered) to facilitate the maximum exertion of force (21).

To summarize the studies reported in this section, the following facts appear to be known with reasonable certainty.

(a) Hand controls are superior to foot controls. There seems to be no reason to prefer wheel over stick control as judged by efficiency of performance in simple tasks. Fore and aft hand motions can be made with slightly greater precision than right and left or rotary hand motions.

(b) Conventional controls should be placed symmetrically with respect to the pilot, and the hands should be at elbow height. No penalty seems to be involved if the pilot adjusts his controls for personal comfort. The shape of controls affects efficiency of performance, and the guiding principle seems to be to shape and place the controls for maximum convenience of grasp.

(c) Full information is not yet available on the accuracy of hand and foot motions of the type used in airplane control. Data are required particularly for various conditions of pressure load. The best present estimate is that linear increments of about 15 per cent may be detected in the linear

displacement of hand operated controls under constant load conditions

(D) Reaction time and rate of motion of controls (1) Reaction time There are many studies that describe the conditions that affect the time required to perceive and respond to a stimulus (17, 28) Reaction time is often measured in laboratory situations that require a minimum of movement, such as may be entailed in pressing or releasing a telegraph key with one finger The basic finding in such studies is that the reaction time is influenced by many variables, among which may be included the sense organ stimulated, the intensity and duration of the stimulus, the motor response involved, the subject's readiness to respond, the complexity of the task, and the subject s age The aircraft designer should know that the shortest reaction time generally reported is of the order of 0 120 sec to sound, 0 140 sec to touch, and 0 165 sec to light These times increase with the complexity of a task, and 0 600 sec is a fair estimate of the time required for such a response as applying brakes to a car after perceiving the cue An early experiment in a cockpit mock up showed that reaction time on a control stick averaged 0 200 sec with a freely moving stick and increased to 0 600 sec with a loaded stick (29) While a simple reaction will usually require about 0 200 sec, a reaction involving discrimination and judgment necessarily will take more time The consequences of such delay may be clear upon reflection that within 0 600 sec an airplane may travel 88 ft while landing at 100 m p h and 733 ft at 500 m p h in the air, and that these speeds are often surpassed at present The effects of such influences as anoxia, fatigue, and drugs, which prolong reaction time, may be examined in Mc Farland s book (20)

(2) Rate of motion of controls Once the response is initiated, the speed of hand motion is a function of the work load and the direction of effort As the stick force per unit displacement increases from 0 to 33 lbs per in, there is a decrease in the rate of stick motion from 75 to 23 in per sec when pulled and 105 to 33 in per sec when pushed (minimum rates of 9 pilots) (3) The rate of push motion exceeds the rate of pull motion One study (29) finds a maximum elevator pull at the rate of 63 in per sec when all conditions of load from 10 to 190 lbs are averaged The slowest rate occurred for two subjects at the maximum load

Flight tests on an F8F 1 airplane (30) show that for approximately equal distances of stick travel (6 to 8 in) the rate of motion dropped markedly from 52 to 10 in per sec as the maximum load increased from 35 to 97 lbs Even though the pilot tried to achieve this motion within 0 200 sec, the actual time for the complete response increased from 0 160 to 0 750 sec as the maximum load increased from 35 to 97 lbs

While the data are admittedly scanty it appears likely that the rate of control stick motion decreases as the load increases Pull rates of the order of 50 in per sec appear reasonable at a load of 35 lbs (maximum elevator limit according to specification) Rates as high as 75 in per sec under lesser loads and as low as 10 in per sec under 100 lb loads may be expected Such evidence as exists suggests that the rate of push motion exceeds the rate of pull motion by about 25 per cent No data on rate of rudder motion are available

Psychological Aspects of Handling Qualities

Stick feel may be observed in terms of the relationship between control stick deflection and control stick force under various conditions, such as with speed and center of gravity position Current specifications on control stick feel are expressed in general terms, which permit considerable latitude in design Thus, it is required that control stick pressure increase with stick deflection from neutral, but the magnitude and regularity of the increase are not specified Control force must also increase with acceleration at a rate of at least 3 lbs per *g* and not more than 8 lbs per *g* It should be clear, however, that 3 lb increments cannot warn the pilot so effectively as 8 lb increments As a matter of fact, the specifications allow such lee

way that the relationship between stick force and stick displacement may be linear or curvilinear The results of the interviews with jet plane pilots and aeronautical engineers show that they believe a linear relationship to be most desirable

Fig 38 6 shows the relationship between aileron deflection and control force at several air speeds in the XP 51 (27), Fig 38 7, for the rudder in the F4U (24) Such

FIGURE 38 6 *Variation of aileron force with total aileron deflection in the cruising condition on the XP 51 airplane (27)*

curves, which are based on flight test data, show that there are families of curves in which control forces increase in a nonlinear fashion with deflection of the control surfaces It may be observed that the curves in the two figures differ from each other in their shape and that this must make a difference in the respective control feel characteristics The following discussion is concerned primarily with two aspects of these curves (1) control displacement versus control force at any air speed and (2) the variation of control forces for any displacement at several air speeds Other gradients, such as between control force and *g*, and between control force per *g* and center of gravity position, also influence the character of the control feel

Consider, first, the relationship between control force and stick position at some particular air speed, which, since we are concerned with fighter airplanes, may be the normal maneuvering speed Fig 38 8 is a conventionalized drawing to demonstrate three possible relationships between displacement and force at one air speed Curve A represents a relationship of the type existing for rudder on the F4U 4 at 353 m p h (24), B, for aileron on the XP 51 at 290 m p h (27), and C, for aileron on the P 47N 1 at 250 m p h (10)

According to curve A in Fig 38 8, initial stick deflections develop large increments

in stick force, while the magnitude of the increments decreases with further stick deflection This is conducive to strong self centering characteristics upon even slight deflection However, the normal work load would be relatively high, and this might lead to unnecessary fatigue Furthermore, since there is no rapid peaking of forces at extreme deflections, there is no warning to the pilot that he may overstress the airplane The shape of this curve is contrary to the nature of human sensitivity

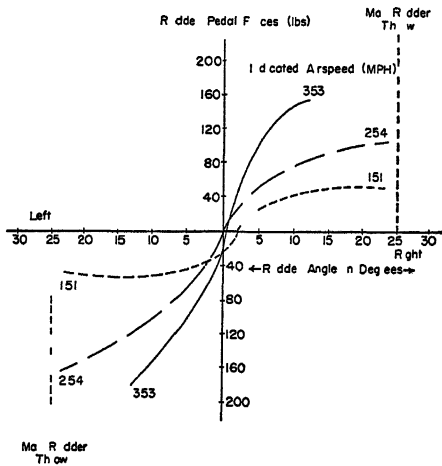



FIGURE 38 7 *Variation of rudder pedal force with total rudder angle in the F4U 4 airplane (24)*

Curve B represents a linear relationship in which stick force is directly proportional to stick deflection over the entire range This is the form often thought to be most desirable, and, indeed, there should be no *a priori* objection to it The deviations from a linear relationship observed in Figs 38 6 and 38 7 are often difficult to avoid because of complex aerodynamic factors In a strictly linear relationship, self centering characteristics may not be strong near neutral, and there may not be a marked warning of an approach to critical conditions

Curve C bears a strong resemblance to the relationship that, as has been shown in this paper, describes the human ability to make discriminations of intensity Since intensity discrimination is a relative and not an absolute ability, the *increasing* amounts of pressure which occur with variation in stick displacement would be experienced as apparently equal steps Thus one might expect curve C to provide maximum sensitivity to differences of

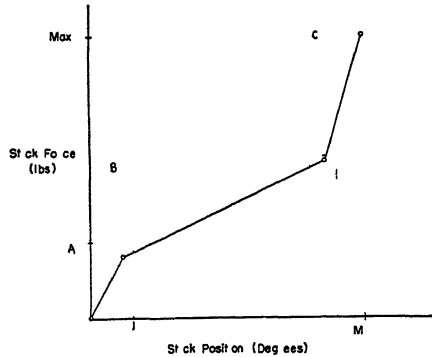



FIGURE 38 8 *Conventionalized curves to demonstrate three possible relationships between stick force and stick displacement at one selected air speed*

pressure This curve might prove deficient at positions near neutral where self centering characteristics would be weak However, control forces would be light over most of the stick deflection range

An ideal force curve should satisfy prob-

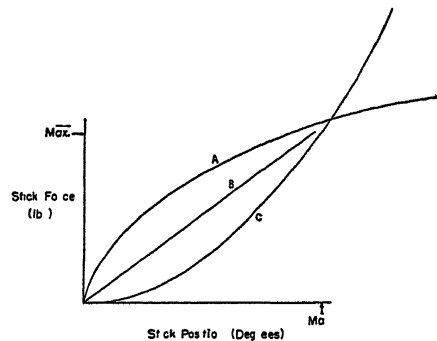



FIGURE 38 9 *Critical areas on a conventionalized stick force versus stick displacement curve for one selected air speed The area designated as A represents the requirements of self centering characteristics the B area represents light stick forces over most of the range and the C area represents warning of extreme conditio*

lems that arise in the three areas identified in Fig 38 9 This is a conventionalized curve and the straight lines, their lengths, and the points of inflection are intended only for purposes of discussion The A band represents the area of initial stick deflection Good stick feel requires that there be strong self centering characteristics, even with slight stick displacement from neutral In practice (as revealed in interviews with pilots) the friction generally inherent in control systems masks self centering and diminishes the feeling of confidence which the pilot gets when the stick is in the groove It is clear that slight stick deflection should produce forces that will exceed the control friction limits permitted by present specifications The amount by which stick force exceeds control friction should be a discriminable magnitude Jenkins (15) has reported that accuracy of performance is poor for stick pressures under 5 lbs (15 lbs for rudder), and this would be a first approximation to an upper limit for the A segment of the curve

The B band represents the area within which most maneuvering occurs There are two major requirements in this area (a) stick forces should be as light as possible to reduce pilot fatigue, and (b) maximum sensitivity of control should be achieved—i e, when constant stick deflection increments produce constant pressure *feel* steps or just noticeable pressure differences The preferred shape of the curve in the B area of Fig 38 9 should be similar to that of C in Fig 38 8 Pressure increments that produce equally noticeable steps are of the order of 10 per cent

Area C in Fig 38 9 represents the problem of extreme stick deflection In this area stick forces should peak rapidly to warn the pilot that he is in danger of exceeding the structural limitations of the airplane This information is transmitted only when the force increments at limiting stick deflections are great enough to be detectable The limit currently imposed by requiring that forces in this area approach the maximum that can be exerted by a pilot is not sufficiently reliable because maximum strength varies among pilots Secondly, present limits may occasionally be exceeded, with unfortunate results, during the emotional stress of combat The general requirements of area C are satisfied by continuing the curve already considered desirable for area B but increasing somewhat the increment ratio, $\Delta I/I$ in the C area A smoothed curve that conforms to the criteria discussed here is illustrated in Fig 38 10 While the opinions of the pilots and engineers who were interviewed cannot be substituted for experimental data, one must report that they all

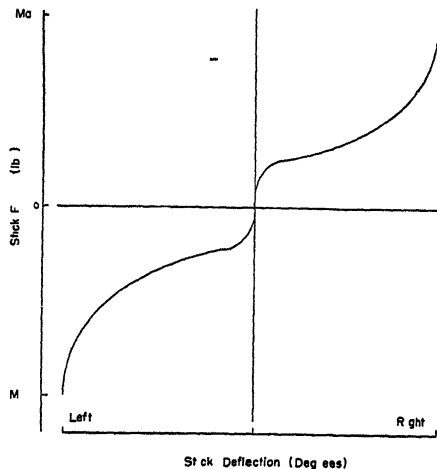



FIGURE 38 10 *Demonstration of a stick force versus stick displacement curve that would satisfy certain conditions proposed in the text This curve would be true for one selected air speed*

agreed, without any reservations, that the stick force curve, as described in Fig 38 10, may prove effective

Since control forces are related to, and increase with, speed, the single curve of Fig 38 10 must be surrounded by a family of curves representing various speeds If control stick feel should yield information on speed (and approach to a stall), these curves must be distinguishable from each other These curves cannot all be psychologically equal The best curve—i e, the one providing the largest number of discriminable pressure steps—should primarily be detailed to the most important tactical requirement In a fighter this might well be the maneuvering speed, while in a trans

port it would probably be the cruising speed

The problem may be illustrated by reference to Fig 38 11, which is a demonstrative plot of the control force at full stick deflection versus air speed The ordinate represents control force (at full stick deflection) with a maximum set by the specifications for stick and rudder The abscissa is related to the flight characteristics of the airplane under consideration otherwise as in curve B The reasoning in this paper would lead to a preference for curve B, the shape of which is dictated by the nature of the human ability to discriminate pressure differences Fig 38 12 demonstrates a possible family of curves for a given airplane showing the relation between stick displacement, required force, and speed The curves are in simplified form because no adjustment is made to allow for the overcoming of initial friction

FIGURE 38 11 *Two possible relationships that may exist between control force at full stick deflection versus air speed*

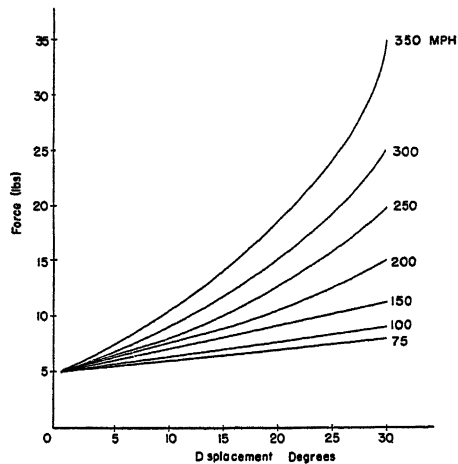



FIGURE 38 12 *Simplified family of curves relating stick displacement stick force and air speed for a particular airplane in accordance with a formula given in the text The maximum air speed is assumed to be 350 m p h maximum control force is 35 lbs and maximum stick deflection is 30° the stalling speed is taken as 75 m p h and the minimal control force is 5 lbs*

One airplane may land at 60 m p h and have a top speed of approximately 180 m p h, thus a factor of 3 (180/60) expresses the range between its lowest and highest speed New airplanes may have a range of 100 to 600 m p h or a speed range factor of 6 Since control forces are limited by an acceptable maximum, as for example 35 lbs in the case of elevator motion, the range between minimum and maximum force must serve various speed ranges In other words, the control force gradient in lbs per m p h (i e, the change in force per unit speed) becomes smaller as the speed range increases This gradient, which is of the order of 0 175 lb per m p h for training airplanes, drops to 0 05 for some planes and has been calculated at 0 03 lb per m p h for some new types The problem confronting the designer is whether this gradient, such as 0 03 lb per m p h, should be spread equally over the speed range as in curve A of Fig 38 11, or

except that all of them start at 5 lbs Curves for other speed ranges may be computed from the following formula,* which was used

$$d = k\ (\log f - \log f_1)\ V_1/V$$

where

d = displacement in degrees
f = control force in lbs
f_1 = minimum force

* This formula was suggested by Dr John D Coakley of Dunlap, Morris and Associates Inc

V = air speed in m p h
V_1 = stall speed
k = a constant defined by setting f d, and V to permissible maximum

These curves are one of several which may be suggested, but before any final curves are adopted, they would have to be validated by flight tests

One important problem is the relation between elevator control force and the weight and balance of the airplane The usual situation is one in which stick force per g decreases as the center of gravity position shifts rearward Another problem is how to ensure the continued effectiveness of the control surface in producing such desired responses as a specified rate of roll, maximum lift coefficient, and directional stability over the entire speed range These are primarily aerodynamic problems, but their solution and standardization would go a long way to simplify such psychological problems as coordination of the controls for smooth flight and consistent flight characteristics for various types of aircraft

It should be pointed out that aircraft control is possible, though not necessarily desirable, without any control stick feel at all An extreme instance is the awkward means by which radio controlled airplanes are flown The pilot operates one or more toggle switches in a bang bang" system, so called because one flick on a switch may cause the airplane to climb while two flicks may cause it to descend There is no feedback of the aerodynamic forces Similarly, the manual adjustments by which maneuvering flight may be accomplished with a gyroscopic autopilot do not supply feedback forces and are different from those required on a normal control stick While flight may be controlled by such means, and contemplated push button schemes promise just this for the future, the real question is whether such methods are adequate for all purposes

The issue may be a minor one for transport type aircraft, where the maneuvering requirement is negligible and where feel may be desirable only for purposes of landing In jet fighters, however, the pilots report an almost complete reliance upon stick feel (and a view of the horizon) during combat maneuvers, with an occasional reference to the Machmeter and yaw strain gage Their experience leads one to the conclusion that some stick feel is highly desirable

The study of control systems should not be limited to conventional forms such as the stick, wheel, and rudder The innovation of booster systems implies that future controls may be of any size or shape and that they may be placed in any location Preliminary investigation should collect information on the various principles of control motion which have been proposed and flight tested, not neglecting those for the prone position One should be careful to guard against the well known tendency to favor those techniques to which one has become accustomed In the event that new control systems may be proposed, the important matters for psychological evaluation are which type (a) permits the most precise flight control, (b) best facilitates learning, and (c) may be operated with the least fatigue In doing so, one should not be deterred by the fact that some of the new proposals may appear to be unconventional An adequate test of various systems, according to their suitability as measured by the human factor, may point the way to significant progress in a field whose main emphasis has often disregarded the very individuals it attempts to serve

References

1 Ashe, W F Roberts, L B, and Bodenman, P *Anthropometric Measurements* U S ASF AMRL, Project No 9, February 1, 1943

2 Bartlett, F C *Instrument Controls and Display—Efficient Human Manipulation* Gt Brit, Flying Personnel Research Comm, Report No 565, December, 1943

3 Beeler, F De, *Maximum Rates of Control Motion Obtained from Ground Tests* N A C A Wartime Report L 100, May 1944

4 Christophersen, D R Kauffman, W M, and Clousing, L A, *Measurements in Flight of the Flying Qualities of a Republic P 47N 1 Airplane* N A C A Memorandum Report for the Air Tech

Serv Comm, AAF, September 24, 1945

5 Craik, K J W, and Vince, M, *Psychological and Physiological Aspects of Control Mechanisms with Special Reference to Tank Gunnery Part I* Gt Brit, Sub committee on Armoured Fighting Vehicles, Military Personnel Research Comm, Medical Research Council B P C 43/254, August, 1943

6 Craik, K J W, and Vince, M, *Psychological and Physiological Aspects of Control Mechanisms with Special Reference to Ground Tank and A A Tank Gunnery, Part II* Gt Brit, Subcomcittee on Armoured Fighting Vehicles, Military Personnel Research Comm, Medical Research Council, B P C 44/322, March, 1944

7 Craik, K J W, and Vince, M, *Psychological and Physiological Aspects of Gun Control Mechanisms Part III Effects of 'Stiffness' and of Spring Centering of Hydraulic Velocity Controls* Gt Brit, Sub committee of Armoured Fighting Vehicles Military Personnel Research Comm Medical Research Council, B P C 45/405, January, 1945

8 Damon, A, *Anthropometric Data on Army Air Forces Flying Personnel* U S AAF ATSC, Exp Eng Section, Memo Report EXP M 49 695 4C, October 3, 1942

9 Gough, M N, and Beard, A P, *Limitations of the Pilot in Applying Forces to Airplane Controls* N A C A T N No 550, January, 1936

10 Grether, W F, *Efficiency of Several Types of Control Movements in the Performance of a Simple Compensatory Pursuit Task,* Chap 17 in Fitts, P M (Ed), Psychological Research on Equipment Design, Washington, D C, U S Government Printing Office, 1947

11 Hick W E, 'Psychological Aspects of Flying Control," *Aeronautics* pp 34–40, July, 1944

12 Hick, W E, *Friction in Manual Controls with Special Reference to Its Effect on Accuracy of Corrective Movements in Conditions Simulating Jolting* Gt Brit, Flying Personnel Research Comm, Report No 623, June, 1945

13 Hick, W E, *The Precision of Incremental Muscular Forces with Special Reference to Manual Control Design* Gt Brit, Flying Personnel Research Comm, Report No 642, August, 1945

14 Honeyman, W M, and Yallop, J M, *Report of an Investigation into the Effects of Asymmetry of Aircraft Controls* Gt Brit, Flying Personnel Research Comm, Report No 621, May 1945

15 Jenkins, W O, *A Psychophysical Investigation of Ability to Reproduce Pressures* Chapt 12 in Fitts, P M (Ed), Psychological Research on Equipment Design, Washington D C, U S Government Printing Office, 1947

16 Johnson, A P, and Milton J L, *An Experimental Comparison of the Accuracy of Sighting and Triggering with Three Types of Gun Sight Handgrip Controls* Chapt 18 in Fitts, P M (Ed), Psychological Research on Equipment Design Washington, D C, U S Government Printing Office 1947

17 Johnson, H M, Reaction Time Measurements Psychological Bulletin, Vol 20, pp 562–589, 1923

18 King, B G, Morrow, D J, and Vollmer, E P, *Cockpit Studies—the Boundaries of the Maximum Area for the Operation of Manual Controls* Project X 651, Report No 3 Naval Medical Research Institute, Bethesda, Md, July 15, 1947

19 McAvoy, W H, *Maximum Forces Applied by Pilots to Wheel Type Controls* N A C A T N No 623, November, 1937

20 McFarland, R A, *Human Factors in Air Transport Design* New York Mc Graw Hill Book Company, Inc, New York, 1946

21 Muller, E A, *Der beste Handgriff und Stiel,* Arbeitsphysiologie, Vol 8, pp 28–32, 1934

22 Patt, D I, *Cockpit Dimensions in Relation to Human Body Size* U S AAF ATSC, Eng Div, Aero Med Lab, TSEAL 3 695 32TT, April 29 1945

23 Patt, D I, and Randall, F E, *Principles of Seating in Fighter Type Aircraft* U S AAF ATSC, Eng Div, Aero Med Lab TSEAL 3 695 58, September 25, 1945

24 Richardson N, *An Hydraulic Booster Equipped Rudder Control System on F4U 4 Bureau No 80770* Report No 7305, October 23, 1946 Eng Dept, Chance Vought Aircraft, Stratford, Conn

25 Vince, M A, *Direction of Movement of Machine Controls* Gt Brit, Flying Personnel Research Comm, Report No 637, August, 1944

26 Warrick, M J, *Direction of Movement in the Use of Control Knobs to Position Visual Indicators* Chapt 9 in Fitts, ˉ M (Ed), Psychological Research on Equipment Design Washington D C, U S Government Printing Office 1947

27 White, M D, Hoover H H, and Garris, H W, *Flying Qualities and Stalling Characteristics of North American XP 51 Airplane (AAF No 41 38)* N A C A Memo Report for the A A F Material Command, April 13 1943

28 Woodworth, R S, *Experimental Psychology* New York Henry Holt and Company 1938

29 *Experiments on the Possible Rate at Which a Pilot Can Pull Back the Control Column in an Aeroplane* Gt Brit Tech Dept of the Advisory Comm for Aeronautics R & M No 282, July, 1916

30 *Report on Flight T 108* (no date) Grumman Aircraft Engineering Corporation Bethpage, L I, N Y

Human Operators and Automatic Machines *

JOHN D COAKLEY

SUMMARY

It is a fairly common belief that automatic production equipment eliminates variability in the product, or at least that such variability as does occur is a purely engineering problem The fallacy of this belief is demonstrated in this study of the production of several operators on the same automatic nylon hosiery machine Costly variability in the hose produced was found not only among operators but also among stockings knitted by the same operator Significant savings in expensive materials and in the cost of subsequent operations could be realized through the reduction of product differences within and among operators on automatic and semi-automatic equipment Product standardization is shown to be a problem to which the psychologist as well as the engineer can offer valuable assistance

DELINEATION OF THE PROBLEM

The fabrication of two objects identical in all respects is possible only in the ideal world In the real world variations in the behavior of workmen, in tool and machine performance and in the raw materials combine to create product variability which seriously affects production costs

Where fabrication requires a certain amount of artistry on the part of the workman, the need for careful psychological study of the operation is readily recognized When, however, articles are produced by automatic machines—machines which presumably do all the direct work of fabricating while the operator merely pushes buttons or pulls levers—the influence of human variation is believed to be eliminated In automatic operations it is supposed that only variations in material and machine performance—mechanical engineering problems—are involved

The manufacture of nylon hose is a semi automatic machine operation in which the operator s function is to pull certain levers as the machine completes each stage of its task On modern machines the operator does not add to or subtract from the number of stitches or from the number of courses knit into a stocking, once the initial setting has been made Superficially, it would seem that the operator could exert little influence on the finished product The variations in the length of the foot and leg which do arise, and which are reflected directly in differences in weight of individual stockings, would seem to be attributable to variations in the nylon yarn and in machine performance

* Reprinted from *Personnel Psychology* Vol 3, No 4, Winter 1950

If the prevalent belief is sound, if automatic machines can produce articles free from the influence of human variation, the production problem is greatly simplified But if the belief is false, no attempt at production control that is based upon it can be highly successful The evidence presented here indicates that the belief is not correct The use of automatic machinery is no guarantee that the articles fabricated will be free from the vagaries of the machine operator The fact that the operator s effect on the finished product is far from obvious increases the need for careful psychological study

The hosiery manufacturer has three very good reasons for wanting to know how operators exert variable influence on the product and how variations can be controlled First, difference in weight means difference in the size of the foot or in the length of the leg or both Such variations give rise to misfitting, variations in wearing qualities, differences in feel and general dissatisfaction on the part of the consumer

Secondly, it has been demonstrated that heavier stockings contain more yarn, and nylon yarn is expensive

Third, stockings made up in pairs must be matched for length before they are packed The greater the variations in weight, the greater the difficulty in matching and the more expensive the packaging operation becomes In fact, the entire finishing operation is slowed down by these variations in hose length

It is clear from these considerations that the isolation and control of human influence on automatic machines is important not only to the manufacturer but to the buyer of his products as well The buyer must pay all the costs of variation

The problem of variations in the weights of finished nylon stockings was studied recently to determine the possible influences of the operators In this investigation a large number of stockings were weighed individually on a specially designed electronic scale Weight was employed because when proper allowance is made for the presence of foreign materials, weight becomes a direct measure of the yarn content of hose The widespread use of nylon, its high degree of standardization, and its physical properties make it especially useful in experimental studies Three different sizes of hose were used—9½, 10, 10½ It would be expected that since size 10 is larger than 9½, the size 10 stocking would require more material and would be heavier Similarly, size 10½ would be expected to be heavier than either of the other sizes The distributions of the weights for the three sizes expressed as percentage deviation from the mean weight of all stockings weighed are shown in Figure 39 1

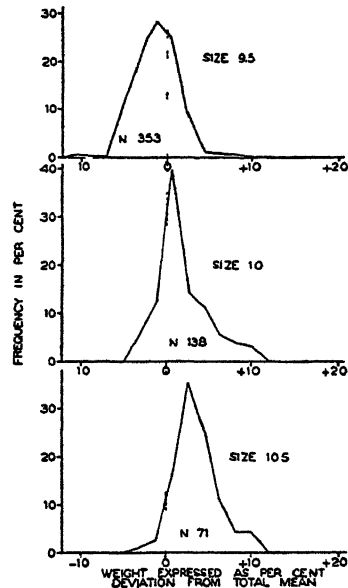



FIGURE 39 1 *Frequency distributions for weights of three sizes of nylon hose*

It will be seen from the distributions of weights that the mean weight increases with stocking size However, the overlapping between sizes is the most notable characteristic of the distributions All of the size 10½ stockings fall within the range of weights obtained for size 10, and more material goes into some 9½ s than goes into most 10½ s Consider the dissatisfaction of a customer who purchases size 10½ stockings and then finds that they are no larger than the average size 9½ Figure 39 1 indicates that all too frequently the

customer actually receives hose which are too small or too large

Upon analysis, numerous causes of these variations were identified, some of them primarily mechanical We will disregard these in the present consideration and de vote our attention to that part of the total variation which is associated with the op erator

Influence of the Operator on the Product

The first step in controlling the oper ator s influence on the product requires that we discover by experimental and sta tistical means the complete range of the operator s influence

If variation is independent of the oper ator, it follows that the same machine, even if run by different operators, should produce the same distributions of weights when it is using yarn from the same lot Three operators, A, B, and C, were asked to operate the same machine, which was adjusted to produce size 9½ stockings The basic adjustments of the machine were unchanged during the study The same lot of yarn was used throughout The number of stitches and courses were the same for all stockings All factors were held constant except the operator and time of day The distributions of weights produced by the three operators in five consecutive 8 hour shifts are shown in Figure 39 2

Comparison of the distributions of A and B shows the weight of A s stockings are uniformly light and vary within a much smaller range than B s Few of A s are as heavy as the average weight for the five runs (shown by the vertical line), while many of B's stockings weigh 10 per cent to 20 per cent more than the average The distribution of C s stockings resembles somewhat those of A, but his total range of variation is closer to that of B

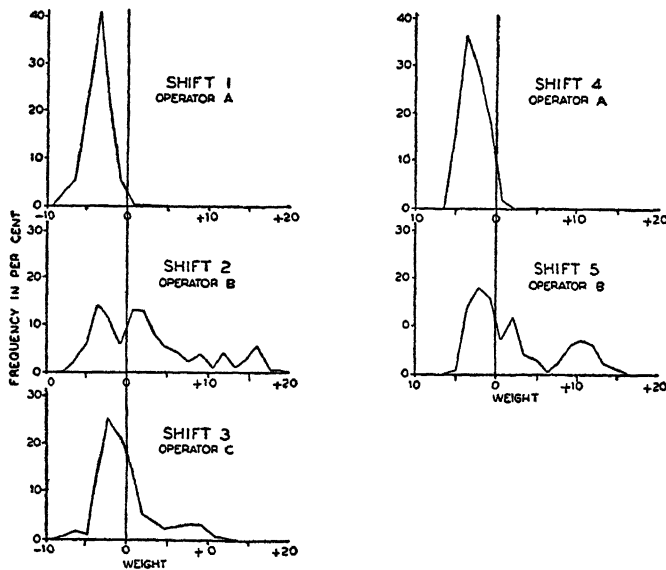



FIGURE 39 2 *Frequency distributions of weights of nylon hose produced by one machine during five consecutive work shifts Total N = 1200 Weights are expressed as per cent deviation from total mean*

Operator A worked shifts 1 and 4, B worked 2 and 5 The weight distributions of stockings produced by each operator on different shifts show marked consistency

TABLE 39 1

Means and Standard Deviations of Weights* of Hose by Day by Man, and Total

Operator	*First Day*		*Second Day*		*First and Second Day Combined*	
	Mean	S D	Mean	S D	Mean	S D
A	204 8	3 4	206 6	3 1	205 6	3 4
B	217 6	12 9	217 0	11 3	217 3	12 1
C	212 0	7 4			212 0	7 4
Total					212 5	10 4

* All weights are in grains

from one day to the next This consistency is so marked that when A s weights for the two shifts and B s weights for the two shifts are each combined the resulting distribution curves retain their characteristic forms This will be seen in Figure 39 3

In the above, we have made comparisons between the weights produced by each operator and the mean weight of all stockings produced during the 5 shifts The mean weights of the stockings produced each are produced during an 8 hour shift Does the operator exert the same influence on every set of 24 stockings or does his influence vary widely from one set to another?

Since 24 stockings are produced simultaneously, it would seem reasonable that the operator s influence must be exerted on all 24 stockings alike Is this a fact or can the operator influence in different ways individual stockings of the same set?

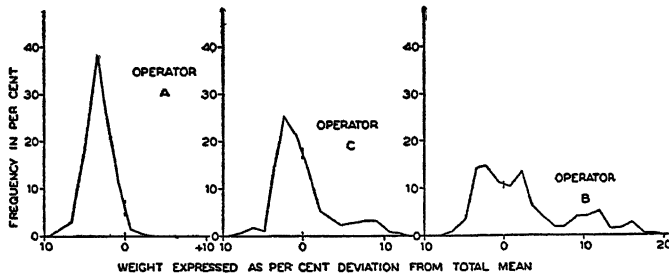



FIGURE 39 3 *Frequency distribution of weights of nylon hose produced by three operators using the same machine Total N = 1200*

during each shift and their standard deviations are shown in Table 39 1 On the average, B s stockings weigh over 13 grains more than A's Not only do B's weights tend to be heavier, but his variation is nearly 4 times as great as that of A and twice that of C

Thus far we have considered differences between operators What about the variability within the product of a single operator? In producing nylon stockings, a 24 section machine produces 24 stockings simultaneously Several sets of 24 stockings

An experiment was designed to answer these questions First, we obtained the average weight of stockings produced in each set of 24 by each of the three operators mentioned above The distribution of these mean weights for sets is shown in Figure 39 4 B produces some sets of 24 stockings whose mean weight is 5 grains lighter than any set produced by A and some that average 35 grains heavier than any produced by A The sets produced by A resemble each other closely, all having a mean weight falling within a range of

15 grains The mean weight of B s sets varies over a range of 55 grains His influence on one set of 24 is very different from his influence on another

If the operator is incapable of exerting influence on individual stockings, it would follow that the scatter of weights of individual stockings in each set of 24 should be about the same The standard deviation

FIGURE 39 4 *Frequency distributions of mean weights of hose in sets of 24 each for three operators Total N = 50 sets*

of each set measures this scatter If the weights of one set are scattered over a wide range and those of another (produced by the same operator) all fall within a narrow margin, it becomes evident that the operator s influence extends not only to the sets as a whole but to individual hose within the set The distributions of the standard deviations by sets for each of the three operators are shown in Figure 39 5 A fair amount of difference in range of variation between individuals of a set is found for operator A while that for B is twice as great

It is clear from these findings that the operator not only causes variations from set to set, making a whole set heavier or lighter, but he also influences individual stockings within the set, making one heavier or lighter than the other

IMPLICATION OF OPERATOR VARIABILITY

With this evidence of operator influences at hand, the behavior of the operators can be investigated in a systematic manner Statistical analysis shows that there are at least 20 different ways in which the oper

FIGURE 39 5 *Frequency distributions of standard deviations for sets of nylon hose produced by three operators A set consists of 24 hose Total N = 50 sets*

ator can influence the weights of stockings In general, the variations are found to arise from relatively simple causes such as the order in which he uses the machine controls, the way he sets the controls, and the way he stretches and inspects hose during knitting

All of these disturbing variables can be controlled readily through proper education and training of the operators In some cases, such as the influence of order of operating controls, it is necessary only to give the knitter an understanding of the manner in which the order of manipulation influences his stockings In some other cases, training procedures are required to standardize and stabilize the most effective behavior patterns

Consider now the manner in which the reduction of variation through training of operators may influence manufacturing costs Of necessity this part of the dis cussion must be somewhat theoretical due to differences in policy in different plants Figure 39 6 is used to illustrate the possible effects of the training program The upper curve of this figure shows the actual dis tribution of weights produced by all three

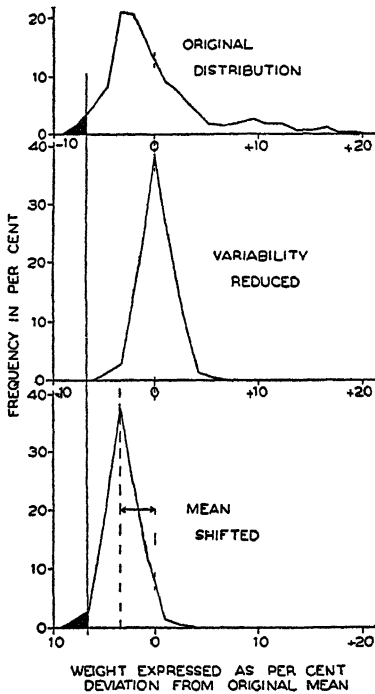



FIGURE 39 6 *Effects of reducing operator variability*

operators combined in this study The vertical dotted line shows the mean of this distribution Assume now that through training we reduce the variable influence of B and C to the same limits found for A Operator A does not, in fact represent an excessively high standard It is known that many knitters produce more uniform hose than he does

If B s and C s variations were reduced to that of A and the mean weight of each operator brought back through adjustment of the machine to the original mean (dotted vertical line) we would have the distribution shown by the middle curve of Figure 39 6 to represent the stockings of all 3 operators This would be the distribution if the manufacturer chose to use the same quantity of yarn that was used prior to training In choosing this mean yarn quantity he frees himself from short stock ings previously rejected (black area in upper figure) This may be the most de sirable course to follow, but there is an other open to him

When human variation is under control, the average weight of the stockings may be altered readily by machine adjustments This mean weight may be shifted up or down at will In the distribution repre sented by the bottom curve of Figure 39 6 the mean weight is shifted to a point where the number of short stockings rejected is equivalent to the number in the top distri bution The mean weight is now about 3⅓ per cent less than the original mean weight Thus the company saves 3⅓ per cent of the total yarn used by the 3 operators Considering the costs of nylon yarn and the quantities used in the course of a year this is no small saving

It is clear from the above considerations that fabrication by automatic machinery, believed previously to present a problem for the production engineer only actually presents a problem for psychologists as well Three functions for the psychologist are revealed

His first function which we are per forming here, is to point out that human behavior does alter the operation of highly automatic machines We have shown that in one situation where variation was at- tributed to raw material and machines the human factors were much more im portant

Secondly, the psychologist is equipped with training and experience especially effective in isolating items of operator be havior which cause variation in the prod uct

Third, the psychologist is equipped to develop effective training procedures through which variability of the operator is controlled

In conclusion it should be clear that the problem of detrimental human influ ence on articles fabricated by automatic

ated object [2] The degree to which peripheral vision is utilized by the reader is an important determinant of perceptual span For instance it is probable that the proficient reader makes maximum use of cues in peripheral vision in reading easy narrative prose Certain factors appear to condition the utilization of peripheral vision in reading

It is obvious from experimental evidence that the extent of the perceptual span is partly determined by the complex central processes of apprehension and assimilation Abundant data show that, as the requirements of comprehension increase, more and longer fixational pauses are required This means a smaller perceptual span

It is possible that another group of factors may also condition the extent of the perceptual span In a group of eye movement studies the authors have demonstrated that there is considerable variation in oculomotor behavior with variations in typographical arrangements This suggests that perceptual span may be conditioned significantly by the physical characteristics of the printed page

In an inconclusive investigation, Luckiesh and Moss studied the influence of type size and line width upon the perceptual span in reading [3] They found a slight decrease in span as the type size was increased from 4 to 10 pt, and a slight increase in span as line width was increased from 13 to 29 picas On the average the span was about 8 5 characters (letters and spaces) They concluded that the number of characters recognized in a typical fixation, *i e* the perceptual span, is substantially independent of type size and of line-width for their readers It is not only desirable to check these results, but to add data for reading text with other typographical variations than type size and line width The purpose of the present investigation is to study the effect of various typographical factors upon the perceptual span in reading

Beginning in 1927, the writers began an extensive series of studies to discover the influence of typographical variations upon speed of reading The results have been summarized in book form [4] Textual material in certain typographical arrangements were read significantly faster than with other arrangements To discover the specific oculomotor patterns responsible for the disclosed differences, a series of eight eye-movement studies were completed [5] The original data of these eight studies have been re analyzed to discover the effect of typographical variation upon the perceptual span

The typographical arrangement employed in each study is described in the tables for each investigation In each comparison 20 college students read 10 paragraphs (29–30 words in each paragraph) from the Chapman Cook Speed of Reading Test, Form A as a standard and 10 different paragraphs from Form B of the same test set in a typographically different arrangement A different group of *S*s was employed for each new comparison

A count was made of the total number of words in each of the two sets of 10 paragraphs of reading material, the total number of picas of linage in each selection

[2] Dodge, *op cit* 33

[3] Matthew Luckiesh and F K Moss, "The Extent of the Perceptual Span in Reading," *Journal of General Psychology* Vol 25, 1941, 267–272

[4] D G Paterson and M A Tinker, *How to Make Type Readable*, 1940 1–209

[5] Paterson and Tinker, "Influence of Line width on Eye Movements" *Journal of Experimental Psychology* Vol 27, 1940, 572–577, "Influence of Line width on Eye Movements for Six point Type," *Journal of Educational Psychology* Vol 33, 1942, 552–555 "Influence of Size of Type on Eye Movements" *Journal of Applied Psychology* Vol 26 1942, 227–230 "Eye Movements in Reading Type Sizes in Optimal Line widths," *Journal of Educational Psychology* Vol 34, 1943 547–551 "Eye Movements in Reading Optimal and Non optimal Typography" *Journal of Experimental Psychology* Vol 34, 1944 80–83, "Influence of Type form on Eye Movements," *Journal of Experimental Psychology* Vol 25, 1939, 528–531, "Eye Movements in Reading a Modern Typeface and Old English," *The American Journal of Psychology* Vol 54, 1941, 113–114 "Eye Movements in Reading Black Print on White Background and Red Print on Dark Green Background" *The American Journal of Psychology* Vol 57, 1944 93–94

(one pica equals about 1/6 in) and the total number of printed characters (letters and letter spaces) in each of the two selections This made it possible to determine the number of words, the number of picas and the number of characters covered in each fixational pause The extent of the perceptual span can be evaluated therefore, not only in number of words read per fixation, but also in terms of the number of picas or characters covered by each fixation Pause duration is included in the reported data to facilitate the interpretations

(1) The results of the first study are given in Table 40 1 The investigation was concerned with perceptual span for all capital printing in comparison with ordinary lower case There were 12 4 per cent more fixations for reading the all capital text The words per fixation were 12 5 per cent less for reading the all capital printing Also 13 6 per cent fewer characters were included in each fixation for reading the all capital text, but 34 4 per cent more distance along a line (picas) was covered per fixation with the all capital material Undoubtedly this increase is due to the fact that for all capital printing, 35 per cent more space is taken for the printing in comparison with lower case text All these differences are significant at or beyond the 1 per cent level when checked by the *t* test for significance of differences In all tables the Prob diff' line refers to the level of significance discovered for the differences

There is another variable involved in the perceptual span that is usually not emphasized when the span is discussed This is the time taken for the fixational pause In this study, for instance the pause duration is significantly less for reading all capital text than for the lower case as shown in column 2 of Table 1 In general this study reveals that type form as a typographical factor affects significantly the perceptual span in reading

(2) The results for the next study are given in Table 40 2 The perceptual span was determined for reading two type faces Cloister Black (Old English) and Scotch Roman More fixations are required for the Cloister Black The span, in terms of words per fixation, picas per fixation, and characters per fixation, is significantly smaller The differences in span are established statistically with a probability at or beyond either the 1 per cent or 2 per cent level For the fixation frequency the change was significant at the 5 per cent level As shown in column 2, Table 40 2, there was no significant change in pause duration The results reveal, therefore, a significant change in perceptual span with change in type face when the type faces are markedly different, *i e* Cloister Black *vs* Scotch Roman

(3) Results obtained in the third study are given in Table 40 3 The perceptual span for reading 10 pt type was compared with span for 6 pt and for 14 pt All line-widths were constant at 19 picas A significantly larger number of fixations was required for reading both text in 6 pt and in 14 pt in comparison with the 10 pt In

TABLE 40 1

Influence of Type Form

Type form	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Lower case	0 23	195 2	1 5	3 2	8 1
All caps	0 21	229 4	1 3	4 3	7 0
Diff	−0 02	+34 2	− 0 2	+ 1 1	− 1 1
% Diff	−8 57	+12 4	−13 3	+34 4	−13 6
Prob diff	01	01	01	01	01

TABLE 40 2

Influence of Type Face

Type face	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Scotch Roman	0 239	210 1	1 4	2 9	7 6
Cloister Black	0 242	220 6	1 3	2 7	7 0
Diff	+0 003	+10 5	−0 1	− 2	−7 9
% Diff	+1 3	+ 5 0	−7 1	−6 9	− 6
Prob diff	30	05	02	01	01

TABLE 40 3

Influence of Size of Type with 19 Pica Line Widths

	Type size	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Study 1	10 pt	0 23	167 7	1 8	3 5	9 0
	6 pt	0 24	180 1	1 7	2 7	8 5
	Diff	+0 01	+12 4	−0 1	− 8	− 5
	% Diff	+7 20	+ 7 4	−5 6	−22 9	−5 6
	Prob diff	01	01	01	01	01
Study 2	10 pt	0 23	152 4	1 9	3 9	10 1
	14 pt	0 21	184 5	1 6	4 9	8 5
	Diff	−0 02	+32 1	− 0 3	+ 1 0	− 1 5
	% Diff	−6 00	+21 1	−15 8	+25 6	−14 9
	Prob diff	01	01	01	01	01

TABLE 40 4

Influence of Size of Type with Optimal Line Widths

	Type size	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Study 1	11 pt	0 22	186 3	1 6	3 1	8 3
	8 pt	0 23	198 8	1 5	2 3	7 8
	Diff	+0 01	+12 5	−0 1	− 8	− 5
	% Diff	+4 2	+ 6 7	−6 3	−25 8	−6 0
	Prob diff	01	02	05	01	05– 10
Study 2	11 pt	0 24	194 4	1 52	3 0	8 1
	6 pt	0 26	197 8	1 51	2 1	8 1
	Diff	+0 02	+ 3 4	−0 01	− 9	+ 01
	% Diff	+7 9	+ 1 8	− 06	−30 0	+ 02
	Prob diff	01	40– 50	70	01	90+

all instances the perceptual span in terms of words per fixation, picas per fixation and characters per fixation was significantly smaller There is an interesting trend in pause duration For the 6 pt print it was significantly greater, probably reflecting increased difficulty in visual discrimination but for the 14 pt, the pause duration was significantly less So size of type, with line width constant, significantly affects perceptual span

(4) The next study was also concerned with type size, but here each size of type was printed in the optimal line width for that type size[6] The perceptual span for reading text in 11 pt type was compared with span in 8 and 6 pt type The results are given in Table 40 4 Examination of the material in the upper part of the table reveals a significant increase in fixation frequency for reading the 8 pt print Although about 26 per cent fewer picas were covered per span, the decreases in words per fixation and characters per fixation are of doubtful significance Pause duration, however, increased significantly The trend for the 6 pt type in comparison with the 11 pt is similar in some respects There is no significant difference in fixation frequency, words per fixation or characters per fixation, but there are significantly fewer picas per fixation and a significantly longer pause duration for the 6 pt In general, therefore, when printed in optimal line widths and leading, changes in size of type do not materially affect perceptual span The amount of linage covered per fixation varies with changes in size of type but the number of words and of characters remains about the same There is, however, a significant increase in pause duration for the smaller sizes of type Thus we find that variation in type size produces significant changes in perceptual span when line width is kept constant, but no significant effects are present when line width is optimal for the particular size of type although in the latter instance pause duration does vary significantly

(5) The effect of varying line width on perceptual span for 10 pt type is shown in Table 40 5 Decreasing line width from 19 to 9 picas produces a significant increase in fixation frequency, a decrease in words and characters per fixation, and fewer picas per fixation When the line width is increased from 19 to 43 picas there is also a significant change in all measures of the perceptual span except words per fixation For both the very short and the very long lines, pause duration is increased significantly These data show, therefore, that varying line width for 10 pt type produces significant changes in perceptual span although the trend is less certain for the very long lines

[6] See footnote 5

TABLE 40 5

Influence of Line Width for 10 Pt Type

	Line width	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Study 1	19 picas	0 22	193 5	1 5	3 1	8 0
	9 picas	0 24	223 9	1 3	2 7	6 9
	Diff	+0 02	+30 5	− 0 2	− 4	− 1 1
	% Diff	+8 1	+15 7	+14 3	−12 9	−13 8
	Prob diff	01	01	01	01	01
Study 2	19 picas	0 22	190 1	1 6	3 2	8 4
	43 picas	0 23	204 9	1 5	3 0	7 8
	Diff	+0 01	+14 8	− 0 1	− 2	− 6
	% Diff	+3 8	+ 7 8	− 6 3	− 6 3	−13 8
	Prob diff	01	01	10	02	02

TABLE 40 6

Influence of Line Width for 6 Pt Type

	Line width	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Study 1	13 picas	0 23	213 0	1 46	2 4	7 3
	5 picas	0 27	235 0	1 27	2 2	6 5
	Diff	+ 0 04	+22 0	− 19	− 2	− 8
	% Diff	+14 3	+10 3	−13 0	−8 3	−11 0
	Prob diff	01	01	01	02– 05	01
Study 2	13 picas	0 23	210 4	1 41	2 32	7 2
	36 picas	0 24	215 9	1 38	2 28	7 3
	Diff	+0 01	+ 5 5	− 03	− 04	+ 1
	% Diff	+3 9	+ 2 6	− 2 1	− 02	+1 4
	Prob diff	01	30	30	30– 40	60

(6) The picture is somewhat different when line width is varied for 6 pt type The data are given in Table 40 6 When the line width is reduced from 13 picas to 5 picas, the span is significantly shortened and the pause duration is significantly lengthened When, however, the line width is increased from 13 to 36 picas no significant change occurs in perceptual span although the pause duration is significantly increased With 6 pt type, therefore, very short lines decrease the perceptual span but lengthening the lines has no effect on the span

(7) The next study is concerned with a comparison of perceptual span for reading an optimal arrangement (10 pt type with 2 pt leading and a 19 pica line width, black print on white eggshell paper) with a non optimal arrangement (6 pt set solid with a 34 pica line width, white print on black background on white enamel paper) The data appear in Table 40 7 In comparison with the optimal arrangement, reading the non optimal print produced significantly more fixations, fewer words and characters per fixation, and 29 per cent less line space (picas) per fixation The increased pause duration was significant only at the 5 per cent level When several deleterious typographical factors are combined in a non optimal printing arrangement, therefore, there is a marked and significant shortening of the perceptual span

(8) The final study is concerned with a comparison of perceptual span for reading black print on white paper with span for reading red print on dark green paper The data are given in Table 40 8 For read

TABLE 40 7

Influence of Type Arrangement

Type arrangement	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Optimal	0 22	191 7	1 5	3 1	8 0
Non optimal	0 24	229 6	1 3	2 2	6 8
Diff	+0 02	+37 9	− 0 2	− 0 9	− 1 2
% Diff	+6 24	+19 8	−13 3	−29 0	−15 0
Prob diff	05	01	01	01	01

TABLE 40 8

Influence of Type Color

Type color	*Mean pause (in sec)*	*Mean total fixation frequency*	*Mean words per fixation*	*Mean picas per fixation*	*Mean characters per fixation*
Black on white	0 23	189 0	1 57	3 1	8 1
Red on green	0 26	235 9	1 27	2 6	6 6
Diff	+ 0 03	+46 9	− 0 3	− 0 5	− 1 5
% Diff	+14 5		+19 1	−16 1	−18 5
Prob diff	01	01	01	01	01

ing the red print on green paper there were significantly more fixations fewer words and characters per fixation, and less linage space (picas) per fixation Also the pause duration was significantly increased For reading the red print on green paper where there was much less brightness contrast between print and paper than for the black on white, therefore, there was a marked reduction in the perceptual span

The implications of these results are clear Certain typographical arrangements reduce the speed of reading The reduction in speed of reading is accompanied by changes in oculomotor patterns Where these changes involve highly significant reductions in fixation frequency, as occurs in most instances, there are significant variations in perceptual span When significant changes in span occur, there may or may not be a significant change in pause duration In general, it may be concluded that typographical variation frequently is an important determinant of perceptual span There are, of course, other factors which influence the span such as the comprehension requirements of the particular reading situation It is quite probable that when such comprehension requirements are exacting, this factor may be more important in determining perceptual span than variation in typography

Summary and Conclusions

(1) The purpose of the present investigation is to study the effect of various typographical factors upon the perceptual span in reading

(2) Data from eight eye movement experiments were analyzed In each investigation the oculomotor behavior in reading a standard was compared with the behavior when some typographical factor such as size of type was varied Measures compared were pause duration fixation frequency, and words picas and characters (letters) per fixation The number of units (words, etc) read per fixation yields the perceptual span A more complete picture of the oculomotor and perceptual changes is obtained by also noting any variation in pause duration

(3) The following typographical changes were found to affect significantly the perceptual span in reading all capital printing vs lower case, Old English (Cloister Black) vs Scotch Roman type face 6 point and 14 point vs 10 point type with line width constant at 19 picas, 9 and 43 pica line widths with 10 point type vs 19 pica line width, 5 pica vs 13 pica line width for 6 point type, a combination of non optimal factors vs optimal typography, and red print on dark green background (low brightness contrast between print and background) vs black on white printing Typographical variations which did not produce significant changes in perceptual span follow 11 point vs 6 point type in optimal line widths, and 13 pica vs 36 pica line widths in 6 point type

(4) Certain typographical variations produced significant changes in pause duration with or without significant changes in perceptual span

(5) The data warrant the conclusion that typographical variation is an important determinant of perceptual span in

reading Optimal typography favors a large perceptual span, and conversely, most non optimal typography reduces significantly the perceptual span

(6) It is conceivable that other factors, such as the requirements of comprehension, may affect the perceptual span more than typographical changes

Differences Among Newspaper Body Types in Readability *

MILES A TINKER and DONALD G PATERSON

So far as the writers are aware no quantitative measures of differences in speed of reading various newspaper type faces have been reported in the literature of journalism or typography[1] This means that newspaper publishers and editors are forced to select a particular type face on the basis of claims set forth by manufacturers and impressions of how one type face looks' as compared with another

The writers, in 1936, made a survey of styles of type faces used in front page body type composition and on editorial pages of 89 newspapers, large and small, published throughout the United States This survey revealed only 15 different type faces in use We selected the following 7 most frequently used type faces for study Ionic No 5, Ideal, Excelsior, Regal No 1, Century Expanded, Textype and Ionic No 2 In addition, two relatively new type faces (Paragon and Opticon) were added Figure 41 1 gives a sample of each type face used

A speed of reading technique was employed to measure the relative readability of the 9 type faces The reading material consisted of Forms A and B of the Chapman Cook Speed of Reading Test In all comparisons Form A was printed in Ionic No 5 type face as a standard Form B, which was read after Form A, was printed in one of the 9 faces In each test form there were thirty items of 30 words each Five items were grouped together in each of 6 paragraphs All material was printed in 7 point solid type, 12 picas wide, on newsprint

Performance on Form B is equivalent to that on Form A on the average However, since there is sometimes a variation from equivalence due to factors other than sampling errors,[2] it is necessary to use a control group

Nine groups of 100 subjects (high school seniors) each were tested In group I, the control group, the typography was identical in Forms A and B (Ionic No 5) Thus a correction can be made in Groups II through IX (the experimental groups) for whatever deviation occurs between Forms A and B of the control group In group II, and each of the successive groups, Ionic No 5 was compared with one of the other types By computing the differences between speed of reading each of the type faces and Ionic No 5, one can discover the relative readability of the 9 type faces

The subjects were tested in small groups Prior to testing, all tests were shuffled to-

* Reprinted from *Journalism Quarterly* Vol 20, No 2, June 1943

[1] This statement is made in spite of claims that visibility meter measurements and blink measurements reflect differences in readability The former measures perceptibility of type at the threshold and the latter is a measure of eyelid reflexes during reading In neither case is there convincing proof available as to their validity as measure of speed of reading

[2] M A Tinker and D G Paterson, Studies of Typographical Factors Influencing Speed of Reading XIII Methodological Considerations, *Journal of Applied Psychology* 1936, Vol 20 132–145 Paterson and Tinker, *How to Make Type Readable* (New York Harper and Bros, 1940)

gether in a systematic manner so that, in any group tested, there were approximately the same number of tests for each type face in the series After the testing was completed, the tests were sorted into the appropriate groups as listed in Table 41 1 Scores were then tabulated and the computations made

The detailed statistical results for the nine comparisons are given in Table 41 1 The entries in this table have been arranged in order from the type face showing the greatest difference from the standard (Ionic No 5) to the one showing the least difference

It will be noted that the three most readable type faces are Opticon, Regal No 1 and Century Expanded Text set in these three type faces are read 6 1 to 7 8 per cent faster than the standard Paragon, Excelsior and Ideal group themselves as being read significantly faster than the standard (4 6 to 5 6 per cent) Ionic No 2 and Textype are read slightly faster than the standard but the differences are not statistically significant

In this study using 900 adult readers, it happens that Ionic No 5, the standard, was read more slowly than any other type face

Standard Ionic No 5

11 When my mother saw the marks of muddy shoes on the floor and all over the nice clean beds, she was surprised to see how careful the children had been. 12 When the little boy

Paragon

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor liv

Control Ionic No 5

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living

Excelsior No 2

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living

Opticon

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living

Ideal

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living

Regal No 1

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living

Ionic No 2

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctoi living

Century Expanded

11 Frank had been expecting a let ter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living in a city

Textype

11 Frank had been expecting a letter from his brother for several days so as soon as he found it on the kitchen table he ate it as quickly as possible 12 A certain doctor living in a city near here

FIGURE 41 1 *Samples of styles of newspaper type (All copy set in 7 point solid, 12 picas wide)*

TABLE 41 1

Newspaper Type Faces

Test Group		Test Form and Type Face	Mean	P E M	Differences between corrected means* in: Paragraphs	Per Cent	r	D/P E diff
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)
I	A	Ionic No 5	14 95	28	0 00	0 0	84	0 00
	B	Ionic No 5	14 43	23				
II	A	Ionic No 5	13 93	24	+1 08	7 8	80	7 49
	B	Opticon	14 49	20				
III	A,	Ionic No 5	14 16	26	+ 90	6 4	86	6 36
	B,	Regal No 1	14 54	24				
IV	A	Ionic No 5	13 51	24	+ 83	6 1	76	5 33
	B	Century Expanded	13 82	20				
V	A,	Ionic No 5	14 40	24	+ 80	5 6	82	5 84
	B	Paragon	14 68	20				
VI	A	Ionic No 5	14 40	27	+ 76	5 3	83	5 11
	B,	Excelsior	14 64	24				
VII	A	Ionic No 5	14 36	26	+ 66	4 6	86	4 96
	B,	Ideal	14 50	23				
VIII	A	Ionic No 5	13 86	21	+ 54	3 9	76	3 79
	B	Ionic No 2	13 88	19				
IX	A,	Ionic No 5	13 92	28	+ 30	2 2	79	1 74
	B,	Textype	13 70	22				

* The differences in column 5 are corrected by the amount of the difference between the mean scores of Form A and Form B of Test Group I, which serves as a control group The correction amounts to 0 52 paragraphs for each test group comparison See footnote 2

NOTE Differences given are for the mean score on Form A minus the corrected mean score on Form B In all test groups Forms A and B were identical typographically, being printed in 7 point solid 12 picas wide, on newsprint Form A in all test groups was printed in Ionic No 5 type face as the standard of comparison Form B varied from test group to test group in style of type face The mean score is the average number of paragraphs of thirty words each in the Chapman Cook Speed of Reading Test (six unit printing arrangement) read in one and three quarter minutes In each test group N = 100 students

The Relative Readability of Newsprint and Book Print *

DONALD G PATERSON and MILES A TINKER

Grateful acknowledgment is given to the Graduate School University of Minnesota for research grant to finance this study

In earlier studies[1, 2, 3] the authors have investigated the readability of newsprint and of book print but no direct comparison has been made between the two kinds of printing There are, however, various hints that newsprint may be read at a slower rate than book print Paterson and Tinker[3] found a consistent tendency for 6 and 8 point book type to be read slower than larger sizes of type The most frequently used type size in newspaper printing is 7 and 8[2] In another kind of study Tinker[4] discovered that in reading 7 point newsprint a greater intensity of light was needed for adequate perception than was necessary with 10 point book type[5] Nevertheless, since newsprint and book print represent somewhat different typographical situations there is not enough evidence for an adequate statement of their relative readability

A direct comparison of the two kinds of printing is made in this study Specifically, the purpose of the investigation is to compare the speed of reading commonly used newsprint and book print

In our survey of newspaper printing[6] the following was the most common practice for body types Ionic type face was most frequently used, with Opticon the most popular of the newer type faces 12 pica line width, 7 and 8 point type and one point leading In the same study we noted that one point leading improves readability of newsprint but that two point gives no added advantage In view of these results and practices we chose the following newspaper typography for use in this study Arrangement number one was 7 point Ionic No 5 in a 12 pica line width with one point leading Arrangement number two consisted of 8 point Opticon in a 12 pica line width with one point leading Both were printed on newsprint paper stock Incidentally, Opticon was the most readable type face of 9 investigated in another study[7] For the book print we chose an optimum typographical arrangement (See Paterson and Tinker, *op cit* 1943) This consisted of Cheltenham type face, 10 point with two point leading in a 20 pica line width on eggshell paper stock Samples of the printing used are shown in Figure 42 1

The reading material consisted of Forms A and B of the Chapman Cook Speed of Reading Test Although performance on Form B is equivalent to that on Form A on the average this is not always true for small samples[8] A control group was intro

---

* Reprinted from *Journal of Applied Psychology* Vol 30, No 5, October 1946

[1] M A Tinker and D G Paterson, Differences Among Newspaper Body Types in Readability, *Journalism Quarterly* 1943, Vol 20, 152–155

[2] M A Tinker and D G Paterson, War Time Changes in Newspaper Printing Practice, *Journalism Quarterly,* 1944, Vol 21 7–11

[3] D G Paterson and M A Tinker *How to Make Type Readable* New York Harper and Brothers 1940, pp 209

[4] M A Tinker Illumination Intensities for Reading Newspaper Type, *Journal of Educational Psychology* 1943 Vol 32, 247–250

[5] M A Tinker, The Effect of Illumination Intensities upon Speed of Perception and upon Fatigue in Reading, *Journal of Educational Psychology* 1939, Vol 30, 561–571

[6] M A Tinker and D G Paterson, *op cit* 1943

[7] M A Tinker and D G Paterson, *op cit* 1943

[8] M A Tinker and D G Paterson, 'Studies of Typographical Factors Influencing Speed of Reading XIII Methodological Considerations *Journal of Applied Psychology* 1936, Vol 20, 132–145

duced, therefore, to check on this equivalence There were 30 paragraphs of 30 words each in each test form The reading time allowed for each form was 1¾ minutes

Three groups of 90 college students each served as subjects In Group I (control) the subjects read book print in Form A and Form B In Group II, Form A was book print and Form B was the 8 point Opticon newsprint And in Group III, Form A was book print, and Form B was the 7 point Ionic No 5 newsprint In addition to the above comparisons, an additional 117 college students ranked samples of the print according to apparent legibility and according to pleasingness In this part of the experiment, samples of 150 words (5 paragraphs of 30 words each) were mounted on cardboard and presented to the readers in a controlled manner

Results and Discussion

Data for the speed of reading comparisons are given in Table 42 1 Results for the control group (Group I) show that a 'correction' must be made by adding 1 59

Cheltenham Book Type
10 point with two point leading

26 James fountain pen went dry when he was doing his homework for school He was very cross because until he got some more glue he could not continue his work 27 The boys saw coming towards them an old woman, bent with sorrow, dressed in deepest black They thought, turning from their play to watch her pass, how happy she looked 28 On

Opticon Newsprint.
8 point with one point leading

26 James' fountain pen went dry when he was doing his homework for school He was very cross because until he got some more glue he could not continue his work 27 The boys saw coming towards them an old woman, bent with sorrow, dressed in deepest black They thought, turning from their play to watch her pass, how happy she looked 28 On

Ionic No. 5 Newsprint:
7 point with one point leading

26 James' fountain pen went dry when he was doing his homework for school He was very cross because until he got some more glue he could not continue his work 27 The boys saw coming towards them an old woman bent with sorrow dressed in deepest black They thought turning from their play to watch her pass how happy she looked 28 On Sunday Mr

FIGURE 42 1 *Samples of book type and newsprint type used in study of relative readability*

TABLE 42 1

Comparison of Speed of Reading Seven and Eight Point Newsprint with Ten Point Book Print

Test Group		*Comparison*	*Mean*	*P E Dist*	*P E Mean*	*Diff Between means in* *Para graphs**	*Diff Between means in* *Per Cent*	*P E Diff*	*r*	*D* / *P E Diff*
(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
I	A	10 pt Cheltenham, 20 pica, 2 pt leading	21 20	2 88	30	0 00	00	00	78	0 00
	B	10 pt Cheltenham, 20 pica, 2 pt leading	19 61	2 63	28					
II	A	10 pt Cheltenham 20 pica, 2 pt leading	21 58	3 13	33	−0 92	4 27	17	86	5 37
	B	8 pt Opticon, 12 pica, 1 pt leading	19 07	2 73	29					
III	A	10 pt Cheltenham, 20 pica, 2 pt leading	21 11	2 59	27	−1 01	4 79	17	79	5 90
	B	7 pt Ionic No 5 12 pica, 1 pt leading	18 51	2 43	25					

* The differences in column 6 are corrected by the amount of the difference between the mean scores of Form A and Form B of Test Group I which serves as a control group The correction amounts to 1 59 paragraphs for each test group comparison

Differences given are for the mean score on Form A 10 pt book type, 20 pica line width with 2 pt leading minus the mean score on Form B printed in newsprint as shown in comparison Book type printed on eggshell and newsprint on newspaper stock In each test group $N = 90$ college students

paragraphs to the mean for Form B Examination of the results for Groups II and III reveals that the 8 point Opticon newsprint was read 0 92 of a paragraph more slowly than the book print, and that the 7 point Ionic newsprint was read more slowly than the book print by 1 01 paragraphs These amount to a retardation in reading rate of 4 3 and 4 8 per cent respectively The critical ratios in Column 10 of the table show that these differences are statistically significant

These results demonstrate that commonly used newsprint even when printed in an optimum arrangement is read much more slowly than book print set in an optimum typographical arrangement

The following factors probably operate to reduce the rate at which the newsprint was read 1 The small size of newsprint type in comparison with the book type makes visual discrimination more difficult, 2 The lower brightness contrast between type and paper for the newsprint would adversely affect discrimination of the printed characters, and 3 Newspaper body types may not be as legible as book type faces It is unlikely however that this third factor is important

Results derived from reader opinions of relative legibility are given in Table 42 2 The order of judgments is 10 point book type ranked first, followed by 8 point newsprint and then 7 point newsprint Note, however, that there is actually very little difference in ranking 8 point newsprint and 10 point book print As has been found before,[9] judgments of legibility do not always agree with actual readability measurements

Readers' opinions of pleasingness are listed in Table 42 3 The order from most to least pleasing is 10 point book type 8 point newsprint and 7 point newsprint Although there is some separation between mean ranks for the book type and the 8 point newsprint, the difference is not great But the 7 point newsprint is considered definitely less pleasing than the others As in the earlier report,[10] pleasingness tends to agree with judged legibility

TABLE 42 2

Book Type and Newsprint Ranked According to 117 Reader Opinions of Relative Legibility

Kind of Type	*Average Rank*	*S D*	*Rank Order*
10 Point Book Type	1 65	79	1
8 Point Newsprint	1 68	58	2
7 Point Newsprint	2 68	60	3

TABLE 42 3

Book Type and Newsprint Ranked According to 117 Reader Opinions of Pleasingness

Kind of Type	*Average Rank*	*S D*	*Rank Order*
10 Point Book Type	1 47	66	1
8 Point Newsprint	1 70	54	2
7 Point Newsprint	2 83	46	3

Summary and Conclusions

1 The purpose of this investigation is to compare the readability of newsprint and book print

[9] M A Tinker and D G Paterson, Reader Preference and Typography" *Journal of Applied Psychology* 1942, Vol 26, 38–40

[10] M A Tinker and D G Paterson, *op cit* 1942

2 Speed of readıng 10 poınt Cheltenham book type was compared wıth speed of readıng 8 poınt Optıcon newsprınt and wıth 7 poınt Ionıc No 5 newsprınt

3 Both kınds of newsprınt were read sıgnıficantly more slowly than the book prınt

4 The slower rate of readıng newsprınt ıs apparently due to the greater dıfficulty of dıscrımınatıng the prınted characters ın comparıson wıth the book type which ıs larger and which ınvolves greater brıghtness contrast between prınt and paper

5 The 10 poınt book prınt and the 8 poınt newsprınt are judged to be about equally legıble but the 7 poınt newsprınt ıs consıdered to be far less legıble

6 The book prınt ıs judged to be most pleasıng the 8 poınt newsprınt next most pleasıng and the 7 poınt newsprınt least pleasıng

Legıbılıty of Numerals The Optımal Ratıo of Heıght to Wıdth of Stroke *

JAMES E KUNTZ and ROBERT B SLEIGHT

Thıs research was supported ın part under the terms of a contract between Specıal Devıces Center, Office of Naval Research and The Johns Hopkıns Unıversıty Contract N5 orı 166, Task Order I The experımental work was done ın the Applıed Psychology Laboratory of Purdue Unıversıty Thıs ıs Report No 166 1 106 Project Desıgnatıon No NR 784 001 The authors gratefully acknowledge the assıstance of John G Gleason who dıd most of the prelımınary work upon the study

The more general factors whıch govern the legıbılıty of numerals may be lısted as follows the sıze of crıtıcal detaıl, contrast between figure and ground, brıghtness contrast between symbol area and background, brıghtness of the symbol, sıze and shape of the surroundıngs, spacıng between symbols, form of configuratıon of the symbol, and the wıdth of stroke or the lıne wıdth of symbol

A careful revıew of the experımental studıes made upon the subject has persuaded the present wrıters that, of all the above condıtıons, the factors whıch stand at present ın greatest need of ınvestıgatıon are the last two ın the lıst, namely, the form of the numeral and ıts lıne wıdth The determınatıon of these prımary factors we have combıned ın the present study by seekıng to establısh the optımal ratıo (*ı e* optımal for readıng) between heıght and wıdth of stroke or lıne (*H/s w*) A secondary determınant ın the experıments has been the dependence of legıbılıty upon brıghtness of the numeral and of ıts background

Apparatus Our maın arrangement was a vıewıng tunnel 8 ft long and 2 ft square A dolly, whıch carrıed the readıng materıal and the tubular lamps to ıllumınate ıt moved back and forth ınsıde the tunnel The posıtıon of the dolly ın the tunnel was controlled by the experımenter (*E*) from hıs posıtıon at one end of the tunnel, whıle *S* bınocularly vıewed the readıng field from the opposıte end *S* s head was fixed by means of a rest The brıghtness of the whıte areas of the field was adjustable by means of swıtches whıch controlled lamps of varıable wattage gıvıng levels of brıghtness of 3, 10, and 31 footlamberts These brıghtnesses were determıned by a MacBeth Illumınometer on whıte surfaces whıch had the same albedo as the whıte of the readıng cards The readıng materıals were photographıc reproductıons of numerals drawn wıth a LeRoy Letterıng Set The *S*s were 14 Unıversıty students, 20–35 yrs of age, of normal acuteness of vısıon

* Reprınted from *The Amerıcan Journal of Psychology,* Vol 63, No 4, October 1950

The stroke-width was obtained by using pen sizes specially machined to give the desired line-widths. The *H/s.w.* ratios were 3.5:1, 4.0:1, 4.5:1, 5.0:1, 5.5:1, 6.0:1, and 6.5:1. The numerals were carefully measured under a microscope. Height was measured from the middle of the horizontal stroke, making the numeral height (*i.e.* not the overall height of the numeral but a midstroke-height), a constant for all numerals at all stroke-widths. Thus a numeral with an overall height of 60 units and stroke-width of 10 units would yield a ratio of 6.0:1 where the midstroke-height would be 5.0:1.

Fourteen reading cards were divided into two groups, half with black numerals on white ground and half white on black. There were 7 different *H/s.w.* ratios. Each card contained two rows of 15 numerals, each numeral appearing three times. Order was assigned by chance, except that no numeral appeared more than twice in a single row and no numeral appeared in immediate sequence. Two sample rows of numerals are shown [above] where *H/s.w.* is 5.0:1.

7 5 6 4 5 3 0 8 6 4 2 9 7 8 1
1 2 9 6 1 3 0 9 5 8 0 7 2 3 1

The reading-cards were presented in a systematic rotation, with ratios, backgrounds and brightnesses counterbalanced. The first distance of presentation was 200 mm. from *S*'s corneas. *S* first read the numerals in order from left to right, and again from right to left. He was instructed to report 'blank' to each numeral not identified. Then the card was moved forward to 180 mm. and again read, but in a new order. This procedure of reading at reduced distances (of 20 mm.) was continued until all numerals were correctly read. The criterion of legibility was the maximal distance of correct reading.

Results. In preliminary trials the brightness of the white parts of the cards was constant throughout at 17 footlamberts. A fairly wide range of *H/s.w.* was used, 4.2:1, 6.0:1, 8.8:1, and 16.0:1. The purpose was to determine an area wherein the optimal ratio might be expected. The results are shown in Fig. 43.1, which gives the average legibility-distance for the various ratios with black-white and white-black. The area that warrants further investigation appears to be in the vicinity of an H/*s.w.* of 6.0:1.

Analysis of the data also revealed a slight but consistent superiority, in terms of legibility distance, for black numerals on white.

In the main experiment a total of 5880 scores, 420 from each of the 14 *S*s, was obtained. The primary treatment of the

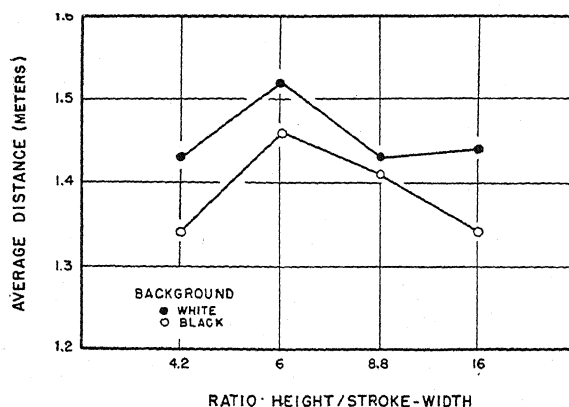



FIGURE 43.1. *Average legibility distance for numerals as a function of height/stroke-width ratio and background (preliminary experiment).*

TABLE 43 1

Analysis of Variance Obtained in a Study of the Optimal Height Stroke Width Ratio of Numerals

Source	*Sum of squares*	*Degree freedom*	*Variance estimate*	*F*	*5%*	*1%*
Subjects (S)	9 448 20	13	726 78	321 58	1 75	2 18
Ratios (R)	127 46	6	21 24	9 39	2 09	2 80
Numerals (N)	42,715 00	9	4,746 11	2 100 01	1 88	2 41
Backgrounds (G)	25	1	25	—*	—	—
Brightnesses (B)	3,400 51	2	1,700 26	752 23	2 99	4 60
Interactions						
$S \times R$	301 47	78	3 86	1 71	1 28	1 41
$S \times N$	2,961 75	117	25 31	11 19	1 24	1 36
$S \times G$	476 71	13	36 67	16 22	1 75	2 18
$S \times B$	3,576 35	26	137 55	60 86	1 52	1 79
$R \times N$	238 25	54	4 41	1 95	1 35	1 52
$R \times G$	14 81	6	2 47	1 09*	2 09	2 80
$R \times B$	101 56	12	8 46	3 74	1 75	2 18
$N \times G$	90 83	9	10 09	4 46	1 88	2 41
$N \times B$	334 10	18	18 56	8 21	1 64	1 99
$G \times B$	179 01	2	89 51	39 61	2 99	4 60
Residual	12,457 74	5,513	2 26			
Total	76,424 00	5 879				

* Not significant

results is the analysis of variance, as sum marized in Table 43 1

The primary variables are the Ss with 13 degrees of freedom, the ratios (R) with 6, the numerals (N) with 9, the back-ground (G) with 1, and the brightnesses (B) with 2 degrees of freedom All inter actions are combinations of the primary variables with the appropriate degrees of freedom

The largest variance is that produced by the numerals (N) This variance is statis tically significant beyond the 1 per cent level of confidence when evaluated against the residual variance Next in order of size are the variances for brightnesses (B), sub-jects (S) and the ratios (R) All of these also exceed the 1 per cent level of confi dence Although the ratios variance is sig-nificant, it still appears to be of minor importance compared to the large size of the numerals and the brightnesses vari ances Nearly all of the double interactions were significant, the one exception being ratio times background ($R \times G$), which failed to reach the 1 per cent level of con fidence The triple interactions were not calculated because they would undoubtedly be small and effects of primary importance were contained in the primary and double interaction variances

The large interactions involve the Ss, the largest appearing between S and brightness ($S \times B$), and suggesting that it makes considerable difference which S read the numerals at which brightnesses This in turn may stand related to a large differential effect of brightness on indi vidual acuity Ss with defective vision are aided by increase in target brightness con-siderably more than others are

The numerals (N) yield the largest main effects variation The interaction be-tween Ss and numerals ($S \times N$) also is larger, but it is small compared to the main (N) variance

With reference to the analysis of vari ance shown in Table 43 1, it may be ob served that, for the range of ratios used in this study (3 5 1 to 6 5 1), the main effect variance, i e, height to stroke width ratios, is significant but relatively small compared to the other main effects variances To be sure, the range of ratios used here is quite limited and large variance would not be expected

In summarizing the information obtained from the analysis of variance it may be

said that, in the main experiment, *S*s, ratios, numerals, and levels of brightness were significant variables (as main effects), but that background (white or black) was not

The optimal ratio of height to stroke width Fig 43 2 shows the average legible distance for black numerals on white back ground and for white numerals on black background for the *H/s w* ratios used in this experiment It will be noticed that if the curves were smoothed they would be almost identical for each background It is also evident that the peak of each curve and 20 per cent are fairly close, due to the parabola at this section of the curve Again, using midstroke height measure ments these percentages would yield *H/s w* ratios of 5 25 1 and 4 0 1

Berger reports optimal ratios of 7 0 1 for black on white and 12 3 1 for white on black when midstroke height measurements are used [2] These findings were based on 4 determinations of the recognizability of each numeral and 4 determinations of their illegibility' by two *S*s using only the numerals 8, 5, 2

Aldrich found width of stroke yielding a

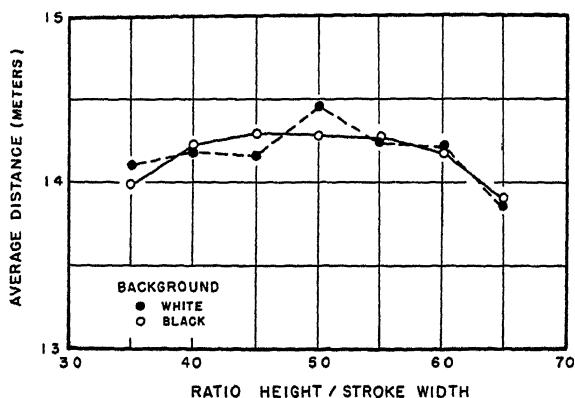



FIGURE 43 2 *Average legibility distance for numerals as a function of height/stroke width ratio and background (main experiment)*

would be at about a ratio of 5 0 1 The peak would not be pronounced, however, and the curves are essentially flat between ratios of 6 0 1 and 4 0 1, indicating negligible differences in legibility for this range of *H/s w*

Our results stand in good agreement with those of Uhlaner, who found that the optimal stroke of three inch block letters is, on the average, closest to 18 per cent of the width or height of the letter' [1] With a midstroke height measurement, this ratio would be 4 55 1 Uhlaner points out in the same study that legibility distances obtained for letters having a stroke of 16 height to stroke width ratio of 7 0 1 was preferable to a stroke width yielding a ratio of 11 0 1, when reading numerals at distances commonly used in reading automobile license plates [3]

Legibility as related to brightness Fig 43 3 shows the average legible distance for numerals of various *H/s w* relative to brightness level under which they were read in our experiments

[1] J E Uhlaner, The Effect of Thickness of Stroke on the Legibility of Letters,' *Proceedings Iowa Academy of Science* Vol 48, 1941, 319–324

[2] C Berger, Stroke width, Form and Horizontal Spacing of Numerals as Determinants of the Threshold of Recognition, *Journal of Applied Psychology* Vol 28, 1944, 208–231

[3] M H Aldrich, Perception and Visibility of Automobile License Plates," *Highway Research Board Proceedings 17th Annual Meeting* Vol 17, 1937 393–412

There is a consistent trend for the legible distance to increase with increase in illumination from 3 to 31 footlamberts The curves show that the optimal ratio remains about the same for all brightness levels, at least for the range of brightnesses here used

Specific numeral legibility That there is considerable variation in the legibility of the various numerals is shown in Fig 43 4 The numeral '1' yields an average legible distance nearly twice as great as the numeral 8 ' There appears to be a consistent tendency for those numerals having most straight lines with least configuration to be most legible Those numerals having most curved lines or most configuration are conversely least legible These considerable variations in specific numeral legibility pose we believe, the most profitable subject for future research in numeral design which looks toward a modification of actual numeral form to increase the legibility of low ranking numerals

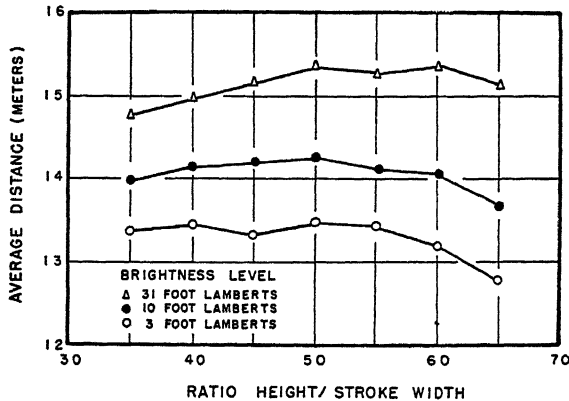



FIGURE 43 3 *Average legibility distance for numerals as a function of height/stroke width ratio and brightness*

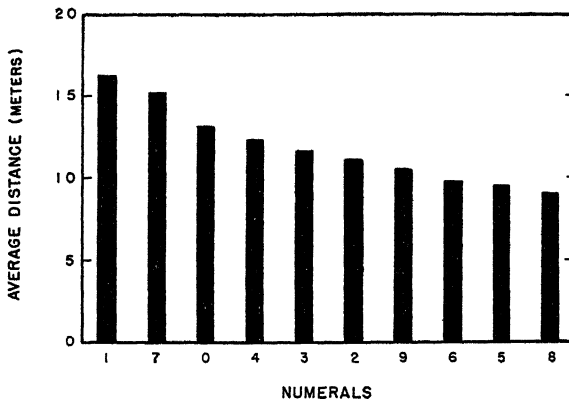



FIGURE 43 4 *Average legibility distance of the various numerals*

Mackworth studying legibility of raid block letters and numbers, modified some letters and numbers and showed that even slight changes in the actual form considerably increased legibility[4] From a com-

[4] N H Mackworth, 'Legibility of Raid Block Letters and Numbers,' *Psychological Laboratory University of Cambridge Eng-*

parison of his new' and 'old designs it is evident that more improvement should be possible

Black on white versus white on black Our results upon legibility for black print on white ground and for white print on black ground are in agreement with some investigators and at odds with others We have found mean legible distances as fol lows black numerals on a white ground, 1 420 m , white numerals on a black ground, 1 419 m , a difference not statis tically significant

Among the investigators who have reported on the black white relationship is Holmes, who found that, in using 10 point type, 'legibility of words printed in black type on white background is 14 7 per cent greater than white type on the black background [5]

Taylor contends that every method of the 4 used by him in reckoning relative legibility of black and white print (6, 8, 10, 12, 14 point type) showed support of the black print on white background [6] He reports in the same study however, that a serifless style (Kabel Lite) was as leg ible as black on white or as white on black, except for the smallest type size used

Ferree and Rand, using broken circles with openings of 1 to 5 as targets reported that for small sizes of objects (1′ or less visual angle) and low intensities, speed of discrimination is higher for black on white than for white on black, but that, for larger sizes of objects and high illumina tions, it is higher for white on black than for black on white ' [7]

Luckiesh, reporting on the use of parallel bars as visual acuity targets, states that "over a large range of brightnesses the black bars on a white background yield sensibly the same result as white bars on a black background [8] This statement seems to conflict with one by Ferree and Rand that ' acuity is found by test to be higher for black letters on white than for white letters on black [9]

From an experiment designed to settle the relative merits of black and white and using 11 point type, Crook found that under the conditions of the experiment, comparing white on black with black on white lettering, there was no significant difference in legibility for the two condi tions [10]

It appears that the most logical conclu sion to draw from the reports noted above is that there may be a statistically signifi cant difference in legibility between the two conditions, but no difference of sig nificant practical importance, except in some special cases

These inconsistent results with regard to legibility of black and white may be of considerable interest because they raise the question of the effects of irradiation, which has conventionally been explained in visual perception as the apparent excess in size of a visual stimulus of relatively high in tensity *e g* of a white figure on a black ground compared with an equal black fig ure on white' [11] It seems important to remember that there may be two different aspects involved in the phenomenon of ir radiation (1) the apparent difference in size when stimuli are well above visual threshold and (2) the threshold of visi bility for figures yielding the same contrast but with opposite figure and ground rela tionships There seems to be a paucity of information concerning the effects of the irradiation in these two situations

Again, if irradiation is to be used as an

---

land, FPRC, Report No 423 (s), April, 1944

[5] G Holmes, The Relative Legibility of Black Print and White Print *Journal of Applied Psychology* Vol 15, 1931, 248–251

[6] C D Taylor, The Relative Legibility of Black and White Print *Journal of Educa tional Psychology* Vol 25, 1943, 561–578

[7] C E Ferree and G Rand ' Intensity of Light and Speed of Vision," *Journal of Experimental Psychology* Vol 13, 1930, 388–422

[8] M Luckiesh, *Light, Vision and Seeing,* 1944 95

[9] Ferree and Rand, *op cit* 394

[10] M N Crook, Further Studies of the Effect of Vibration and Other Factors on Legibility of Numerals, AMC, Wright Field, Dayton Ohio, Report No TSEAA 694 1K, 21 Oct, 1947

[11] H C Warren, *Dictionary of Psychology,* 1934, 144

explanatory concept for certain problems in legibility, then the basic problem of figure ground relationship may be of considerable importance This is illustrated in the following remark by Ferree and Rand[12]

Near the threshold of acuity, i e for small visual angles and low intensities, speed of discrimination is higher for such objects as a black letter on white (detail to be discriminated white) than for the white letter on black (detail to be discriminated black) Because of irradiation in the image formed on the retina, the detail to be discriminated in case of the black letter on white is larger than that to be discriminated in case of the white letter on black

It is obvious that this explanation becomes untenable if a contrary point of view is accepted, i e, that with a black figure on white the detail to be discriminated is *black* The precise effect of irradiation and the conditions under which it is effective would seem to be a fair problem for future research

[12] Ferree and Rand, *op cit* 419

Summary and Conclusions

A study was made with numerals of varying height and stroke width read at various distances in order to determine the ratio yielding highest legibility Effects of brightness and background on legibility were likewise studied The following conclusions seem to be warranted

(1) Using a midstroke height measurement for the numerals made with a modified LeRoy Lettering Set, the optimal height *vs* stroke width ratio (*H/s w*) s between 6 0 1 and 4 0 1, or, more precisely stated, at 5 0 1

(2) The optimal ratio is approximately the same for numeral brightnesses of 3 0, 10 0, and 31 0 footlamberts

(3) The several numerals used show clear differences in legibility They rank from most to least legible in the order 1, 7, 0, 4, 3 2, 9, 6, 5, 8 indicating a need for modification in the form of certain numerals in order to improve the legibility of those of low rank

(4) Legibility of black numerals on white background and that of white numerals on black background is the same under the conditions of this experiment

II Stroke-width, Form and Horizontal Spacing of Numerals as Determinants of the Threshold of Recognition *

CURT BERGER

Night Experiments [1]

Recognizability of white numerals with reflected light and luminous numerals during night conditions First, the optimal light intensities for white numbers of optimal day vision under night conditions were determined, using 5, 10 and 15 watt lamps as light sources for reflecting light The exact location of the light source was as generally used in Denmark, in the cen

* Reprinted from *Journal of Applied Psychology* Vol 28, No 4 August 1944

[1] The Danish as well as many other license plates are lighted from below during the night, and this light source is also used partly for the red rear light Experiments have shown that the use of this red rear-light so close to the number diminishes its legibility between 10 per cent to 20 per cent It must therefore be recommended that the red rear light be used as far away from the number plates as possible New motorcar models already have the red rear-light on one or both outer sides of the rear and the license plate alone in the middle of the rear In that case the red rear light does not affect the legibility of the numerals on the license plate any more

TABLE 44 1

Threshold Recognition of White Numbers, Illuminated from the Side of the Observer (il) and of Luminous Numbers (lu) Illuminated from behind Il = 15 Watt, lu = 10 Watt Five observers and 5 different widths of strokes are used The average represents 40 single threshold measurements (Number 8)

O Sch *m*		*O Ha* *m*		*O Ra* *m*		*O Sch* *m*		*O Gi* *m*		*Average* *m*	
il	lu	il	lu	il	lu	il	lu	il	lu	il	lu
2 mm =25 9	1 mm =27 4	2 mm =39 1	1 mm =40 6	2 mm =29 8	1 mm =31 9	2 mm =27 1	1 mm =32 6	2 mm =28 5	1 mm =37 6	2 mm =30 1	1 mm =34 0
4 mm =24 8	2 mm =27 8	4 mm =39 9	2 mm =41 9	4 mm =30 9	2 mm =33 0	4 mm =29 1	2 mm =28 6	4 mm =27 0	2 mm =38 8	4 mm =30 3	2 mm =34 0
6 mm =23 8	3 mm = 26 9	6 mm =39 4	3 mm =35 4	6 mm =31 1	3 mm =32 2	6 mm =29 1	3 mm =26 1	6 mm =31 3	3 mm =36 8	6 mm =30 9	3 mm =31 5
8 mm =22 0	4 mm =24 9	8 mm =39 1	4 mm =34 1	8 mm =31 5	4 mm =30 7	8 mm =28 7	4 mm =24 1	8 mm =26 3	4 mm =34 1	8 mm =29 5	4 mm =29 5
10 mm =22 1	5 mm =21 8	10 mm =37 9	5 mm =32 1	10 mm =31 1	5 mm =29 8	10 mm =27 9	5 mm =23 2	10 mm =25 8	5 mm =33 6	10 mm =28 9	5 mm =28 1

TABLE 44 2

Threshold Recognition of White Numbers, Illuminated from the Side of the Observer (il) and of Luminous Numbers (lu) Illuminated from behind Il = 15 Watt, lu = 10 Watt Four observers and five different widths of strokes were used The average represents 32 single threshold measurements The number 2 was used with average night conditions

O Ha *m*		*O Schoe* *m*		*O Ra* *m*		*O Gi* *m*		*Average* *m*	
il	lu	il	lu	il	lu	il	lu	il	lu
2 mm =26 9	1 mm =35 3	2 mm =30 3	1 mm =32 3	2 mm =25 0	1 mm = 33 1	2 mm =23 3	1 mm =32 3	2 mm =26 4	1 mm =33 2
4 mm =27 0	2 mm =35 3	4 mm =31 3	2 mm =32 3	4 mm =27 1	2 mm =33 4	4 mm =36 2	2 mm =31 3	4 mm =27 9	2 mm =33 1
6 mm =25 9	3 mm =32 6	6 mm =31 4	3 mm =33 6	6 mm =33 6	3 mm =31 3	6 mm =27 3	3 mm =28 9	6 mm =28 0	3 mm =31 5
8 mm =25 2	4 mm =29 9	8 mm =32 6	4 mm =31 6	8 mm =27 9	4 mm =29 8	8 mm =26 8	4 mm =27 6	8 mm =28 1	4 mm =29 7
10 mm =23 7	5 mm =27 8	10 mm =30 7	5 mm =28 3	10 mm =27 3	5 mm =28 9	10 mm =25 7	5 mm =26 2	10 mm =26 8	5 mm =27 8

ter, some inches below and before the illuminated license plate Ten to 15 Watt lamps were found practically equal and better than 5 Watt A further increase of light intensities did not seem to improve recognizability further

Similar experiments were made with luminous numbers of equal appearance, cut out of card board and illuminated from behind, the light source being in a light tight box Ten Watt lamps were found most satisfactory

These experiments were made only in order to determine roughly the optimal average light intensities for reflected as well as luminous numerals under the above described night conditions Actually the experiments described in this paragraph should have been made in combination with these preliminary experiments, since probably each particular intensity used will show a maximum recognizability with one particular width of stroke, changing into a slightly more slender stroke when intensity increases if reflected light is used In the case of luminous numerals such changes are probably limited to a very narrow range of extremely slender stroke widths in combination with relatively very great changes of intensities which are not practicable Such a thorough investigation would have been a task of considerable proportions Nevertheless, it would have been carried out but for two reasons

First if reflected light is used under night conditions, the numerals were bound to have the width of 6 mm, namely the optimal daylight structure since it is more or less practically impossible to use on the same plate two differently constructed numbers for day vision and night vision, both with reflected light

Second, the intensities used should be kept low because of practical reasons, that is that the lamps are run from motorcar batteries Furthermore, it could be concluded from preceding experimental results of a theoretical nature, that using night conditions under which the human eye, partly or totally dark adapted, has greatly increased light sensitivity, intensities would rather have to be decreased than increased in order to obtain better recognizability The lower the intensities, the less aberration and the fewer glaring effects at night The main improvement was to be expected from investigating differences between reflected light conditions and luminous numerals, illuminated from behind The preliminary study of optimal light intensities was therefore limited to the above described rough determination of optimal average light intensities

Subsequently, experiments were made with the number 8 and the number 2, using widths of strokes between 2 and 10 mm for numerals with reflected light and widths of strokes between 1 and 5 mm for luminous numerals The results are shown in Tables 44 1 and 44 2 and the average of all these experiments in the curves of Figure 44 1

The very slender numerals with a width of stroke of 1 to 2 mm illuminated from behind (luminous), are on an average (5 subjects used) 10 per cent to 18 per cent more recognizable at the above mentioned intensity than the white numerals of optimal stroke-width with reflected, optimal illumination Furthermore, the wider the stroke of luminous numerals (beyond 2 mm), the less recognizability was obtained, while with numerals using reflected light, the recognizability increases with the width of the strokes until a maximum is reached between 6 and 8 mm, under the above described illumination conditions decreasing again with wider strokes

In Table 44 3 results are compared directly between luminous five number groups with only 2 mm stroke width, the old Danish numerals with reflected light and optimally legible white numerals of 6 mm stroke width and with reflected light *The luminous slender numerals are on an average about 37 1 per cent more recognizable than the Danish and about 17 8 per cent more recognizable than the optimal white numerals of 6 mm stroke width with reflected light*

Discussion of the Results

The results obtained in this investigation are interesting in several respects

First, the previously mentioned work of

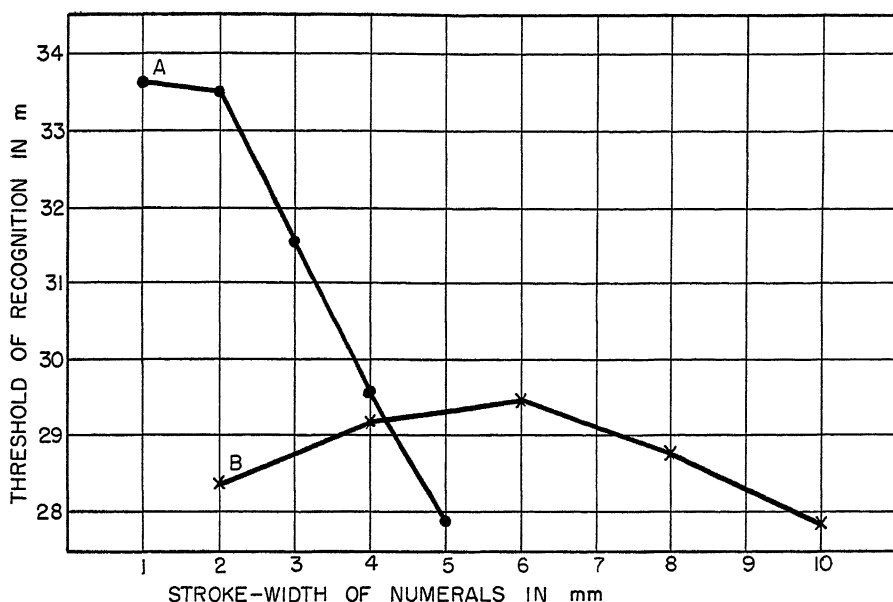



FIGURE 44 1 *Dependency of the threshold of recognition upon stroke width of numerals during night conditions (partly dark adapted) Curve A = luminous numerals Curve B = white numerals with reflected light*

TABLE 44 3

Comparison of Threshold Recognition between 2 Five Number Groups during Night Conditions The Danish Numbers as Used before the War, the New Daylight Numbers with Reflected Light (15 Watt) and Luminous Numbers with a Stroke Width of 1 5 mm, Illuminated from behind with 4 × 10–40 Watt Four Observers were used

Ob server	*Number Group*	*Danish Number m*	*New Numbers with Reflected Light m*	*Luminous Numbers, m*	*Luminous Numbers Improvement in m*	*Luminous Numbers Improvement in per cent (compared with Danish num)*
Gi	80 236	25 3	29 4	34 7	9 4	37 6
	91 475	24 0	30 5	35 8	11 8	49 1
Schoe	80 236	26 5	29 7	35 3	8 8	32 6
	91 475	25 3	32 4	34 0	8 7	34 8
Scha	80 236	21 6	23 8	28 2	6 6	30 0
	91 475	22 8	26 4	30 0	7 2	31 3
Ma	80 236	23 7	26 2	34 2	10 5	43 7
	91 475	24 1	26 8	33 2	9 1	37 9

K Dunlap [2] can be thoroughly analyzed Since it is the only exhaustive experimental study of the legibility of numerals on license plates known to the author it seems worth while to draw clear conclusions from this study with respect to Dunlap s statements

[2] K Dunlap, *Report Highway Research Board* National Research Council, Division Office, 1932, App E, p 3, article 4

(a) *Light background with dark numbers gave best results* The validity of this general statement is due and limited to the particular structure of the numbers, light on a dark background, and dark on a light background compared As a generalization it is definitely wrong The actual comparison should be made between optimally recognizable white numbers (relatively slender) and optimally recognizable black numbers (with a wider stroke) Any other comparison is a chance comparison and will depend upon the fixed absolute proportions between stroke width and breadth and height of the numerals chosen If the comparison is made between optimally recognizable white and black numerals, it has been shown above that the white numerals on a black background are about 8 8 per cent more recognizable than the black on white, if used singly The question becomes somewhat more complicated, if 5-number groups are used It is probable that black numerals will require smaller inner' and 'outer distances for adjustment to standard than the white numerals But since the stroke-width for black numerals optimally recognizable, is so much larger than for white numerals, it is doubtful how much the loss of recognizability for the single numbers can be compensated by less spacing between them or inside them As long as the height of the numerals and the standard' area are kept constant, it is probable that the optimum recognizability obtainable for black numerals by adjusting their inner and outer distances might be close to equality to optimally recognizable white numbers, but not more

(b) *Plates without borders seem best* This problem as described above, depends largely upon two factors whether black numerals on a white background are used or vice versa and upon the width of the borders and their spacing away from the top and bottom as well as the sides of the numerals, if white numbers are used Contrary to Dunlap's statement, our above described experiments show that making a white frame of a width equal to the numeral stroke width and spacing it at the correct distance increases the legibility of white numerals considerably Whether a similar effect is exercised by a particular size of the white background or a black border (contrast) for black numerals, has to be decided by further experiments

(c) *Numerals spaced further apart gave highest legibility* It is true that at present most numerals are spaced too close together Thus Dunlap found those numerals more legible which were spaced farther apart Nevertheless, as shown above, there is no reason to make the spacing between the numbers larger than necessary It should correspond to the recognizability of

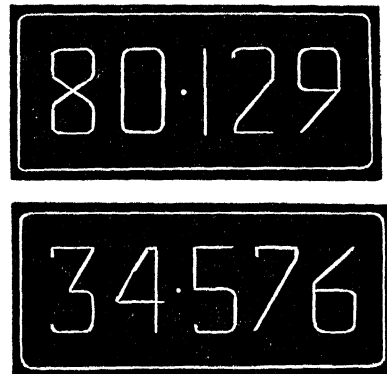



FIGURE 44 2 *Final appearance of all numbers luminous on a black background for night vision reduced to 1/63rd of the original area (Compare Tables 44 1 and 44 2)*

the single and standard number, which can be found by the procedure described above, and which is different for a five number group than for two numbers alone If the space between the numbers is larger than that corresponding to standard, a further increase cannot possibly give 'higher" recognizability, because the numbers themselves are already illegible at a distance from the eye, in which the numerals would fuse into each other It also would enlarge the license plates unduly Dunlap s statement thus is true only until a definite limit is reached

(*d*) *Plates in which the numbers did not exceed 25 per cent of the total area of plate were most efficient* This is not correct for white numerals, where the background and its size has no influence upon

the recognizability of the numerals, being black But even if we consider as total area" of the plate for white numerals the area covered or surrounded by the white frame, the white numbers, found in this investigation as optimal, cover about half of that area It is also doubtful whether this statement is correct for optimally con structed numbers, even if they are black Nevertheless in the case of a white or colored background, its size might have a particular surrounding effect which has to be investigated for each case separately

(*e*) *Numerals with slender stroke were more efficient* It is true that in most cases, actually used in practice, the strokes are too wide But there is a definite limit to the slenderness of the strokes, both for white numbers on a black background, and particularly for black numerals on a white background with daylight It is different for both cases but if the numerals are chosen more slender than corresponding to optimal standard recognizability, the legibility will diminish rapidly

Besides our wide amplification of Dun lap s results it is a notable discovery made up on the basis of previous theoretical work, that recognizability depends upon form as well as upon character and color, black, white, luminous and so on of the numbers used and not exclusively upon the equality of details as assumed by Snellen and others The difference between the results obtained in this investigation and Snellen s principle is shown in Figure 44 3 for white numerals with daylight Two particularly important conclusions can be drawn from this discovery (1) Since the underlying theory was based upon experi ments made under a moderate degree of light adaptation, while for our night con ditions partly dark adapted observers were used, visual resolution of luminous details for the partly dark adapted eye might not differ basically from resolution of the mod erately light adapted fovea (2) If we take the ' inner distances" of the numerals re quired for standard recognizability, as the basis for an estimation of the ' goodness" of their forms with daylight, they range as follows (starting with the best), 2, 0, 7, 3, 5, 6, 4 and 8 The zero, closest to the circle, is not the best, while the first 3 are less structured than the following three numerals, which again are simpler than the numerals 4 and 8 Thus it can be con cluded, that the more complicated the structure of a numeral, of equal stroke width and under otherwise constant con ditions, the more difficult is its recogniz ability Neither Gestalt assumptions[3] nor the work of Helson and Fehrer[4] are sup ported by our results Form plays a defi nite role for the recognizability of nu merals, but its effect must rather be ex plained by structural and functional prop erties of the retina or a corresponding organization of central parts of the nervous system The angles formed by connecting lines between horizontal or vertical end points which are optimal for the recogniz ability of the forms, seem to be independ ent of configuration, showing the same tendency for all numerals These results lead to the conclusion that the recogniz ability of numerals is, in spite of its being a very complex function even under the simplest life conditions, directly related to the size as well as the function of the structural basis of the human eye, namely the single functional units of the retina[5]

Finally, we must mention a point of view which, though not predictable, opens up an interesting perspective The nu merals, presented as final results for day light in this study have been de veloped exclusively on the basis of op timal recognizability using the psycho physiological method of thresholds But they seem also to possess high esthetic quality Although the appearance of the numerals when near at hand is not directly comparable to their appearance near threshold distance, there may be a relation between recognizability and esthetic qual

---

[3] K Koffka, ' Psychologie du optischen Wahrnehmung, *Bethe s Handb norm u Pathol Physiol* 1931, Vol 12, 1215–1271

[4] H Helson and E V Fehrer, The Role of Form in Perception," *American Journal of Psychology* 1932, Vol 44, 79–102

[5] C Berger and F Buchthal 'Der Einfluss von Belenchtung und Ausdehnung des ger eizten Netz Lautoreales sowie von Pupillen durchmesser auf das Auflosungsvermogen des linmetropen Auges,' *Shand Arch f Physiol,* 1935, Vol 71, 173–199

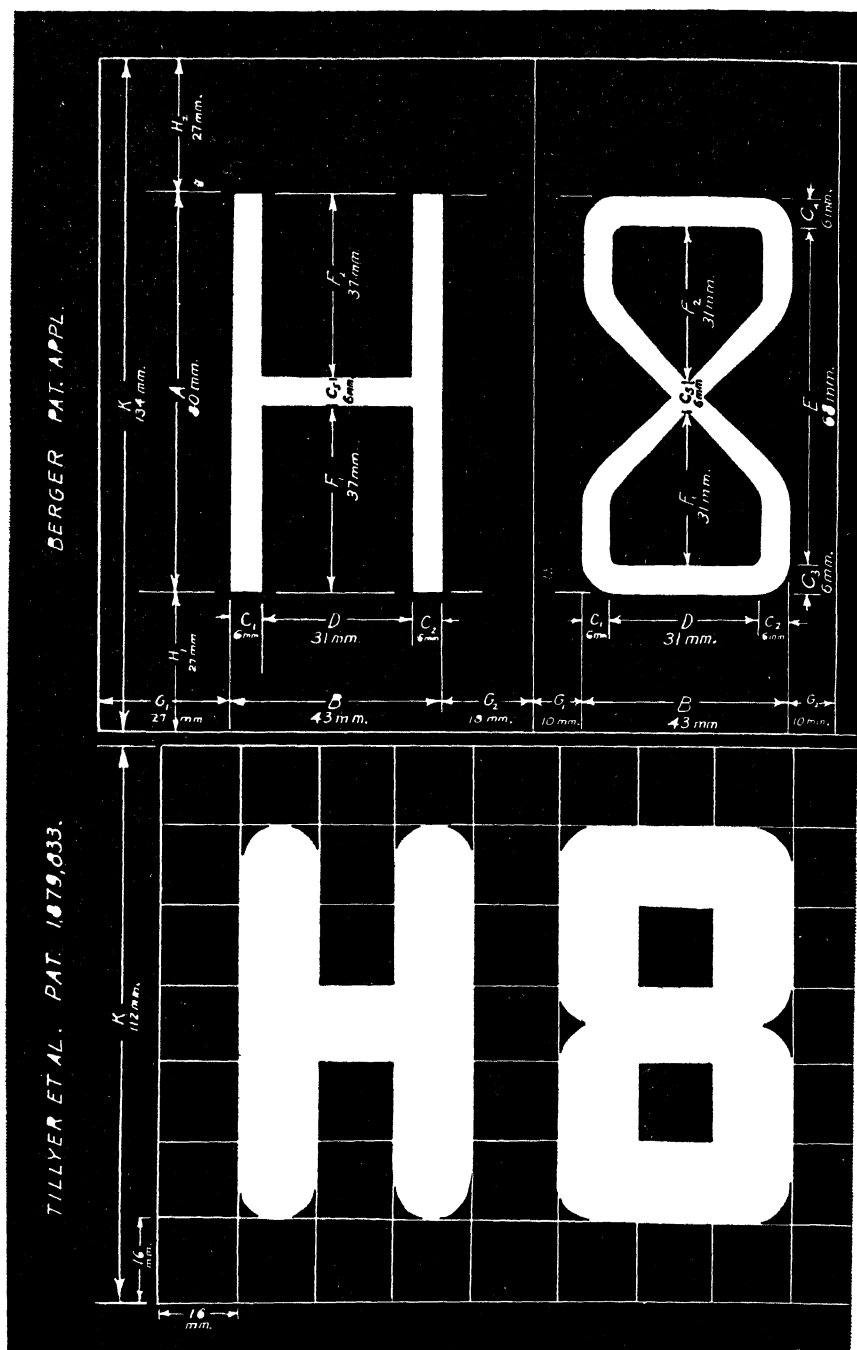



FIGURE 44 3 *Comparative representation of the 'norm" number 8 and a letter H and the same symbols constructed on Snellin's principle (Note that inner distance (D) of letter H is not adjusted to standard)*

ity [6] Symbols of optimal recognizability may prove, under definite conditions, to be the symbols of highest esthetical appeal under the same conditions, even when regarded at shorter distances than the threshold The proof of this assumption must, however, be undertaken in a separate investigation If found correct, the structure and properties of the receptive visual organ, particularly of the retina, might prove not only, as recently suggested by Polyak,[7] to be the real basis of the *a priori* qualities of space (Kant), but also for the esthetic qualities of seen objects Our experiments indicate that the solution of this problem is complicated by a very great number of factors in ordinary vision It has to be specifically approached for each particular situation but a solution might be obtained for certain definite conditions Our results suggest that the properties, structural and functional, of the human retina offer a more direct way of approaching this problem than do the configurational theories

SUMMARY

1 A method and procedure is described to investigate the influence of stroke width of white and black numerals, specific form factors, distances between the strokes of numerals, distances between two numerals and surroundings upon the threshold of recognition, with a view towards improvement of the recognizability of the numerals The same method is applied to conditions of night vision of medium dark-adaptation

2 A construction is found for 9 numerals, white on black background, luminous during night conditions, which are optimally recognizable (standard area 42 mm × 80 mm) and which at the same time are adjusted to standard in such a way, that each single numeral as well as 2 and 5 number constellations appear or disappear at the same distance from the eye, a distance which is the greatest possible distance for the particular area chosen

3 During the investigation the following results were registered

(*a*) Numerals, white on a black background, have an optimal average recognizability, if their stroke width is 6 mm on an area 42 mm × 80 mm Numerals, black on a white background, have an optimal average recognizability, if their stroke width is 10 mm on an equal area The proportion of this stroke width to the inner horizontal distances of the numbers is in the first case about 1 5, in the second about 1 2 2 The white numerals are under these conditions, singly about 8 2 per cent more recognizable than the optimally constructed black numerals for the same area

(*b*) Investigating many detailed characteristics of form, it was found that the angle under which two horizontal lines are connected (or two vertical lines) has a particular importance for the recognition of a numeral The recognizability is best with angles which cut the adjacent parts of the area into two equal halves A vertical or horizontal connection is least recognizable

(*c*) A 5 number group requires about 10 per cent more space between the numerals than 2 numbers alone, even though the space between the 2 numerals already has been adjusted to standard

(*d*) A white frame around white numerals improves their legibility only if at a certain distance from the top and base of the numerals, and only if its width is equal to the stroke width of the numerals Then the improvement is about 9 per cent

(*e*) Under ordinary night conditions (medium dark adaptation) very slender, luminous numerals at threshold brightness are about 17 8 per cent more recognizable than optimally constructed white numerals with reflected light

(*f*) The relations of these results to previous theoretical work are discussed, and it is pointed out that they not only corroborate theoretical conclusions concerning structural as well as functional concepts of the human eye, particularly its fovea, but may also lead to interrelations between purely physiological and certain esthetic aspects of human vision

[6] R S Woodworth, *Experimental Psychology* New York Henry Holt & Company, 1938, 368–392

[7] S L Polyak, *The Retina* Chicago The University of Chicago Press, 1941, 444–446

Approach Speeds and Changes in Sign Size and Location on the Highway *

C H LAWSHE, JR

This study is a portion of a thesis submitted to the Faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy June, 1940 Acknowledgment is due Dr Joseph Tiffin who directed the research, the members of the State Highway Commission of Indiana and the officers and advisory board members of the Joint Highway Research Project whose financial support and cooperation made the investigation possible

In comparing the relative merits of traffic signs on the highway, whether or not the driver can read or identify them is less important than his response in terms of the behavior that is expected or desired Since numerous signs now in use have as one of their functions the reduction of speed at some point where higher speeds are thought to be hazardous, it seems that the speeds of drivers in the presence of these signs should be taken into account in any evaluation of the effectiveness of the signs The present investigation has as its purpose the establishment of a technique utilizing speed measurement in the evaluation of certain traffic signs and signing practices on the highway

Experimental procedure The multiple-speed recorder described by Lawshe [1] was employed to measure the speeds of motorists without their knowledge as they approached an intersection but at a point where preliminary investigation had shown that deceleration for the intersection had not yet begun Likewise, the speeds of these same drivers were also recorded nearer the intersection at a point on the highway where the driver had begun to reduce his speed These two points at which speeds were recorded were approximately 1100 feet and 400 feet respectively from the center line of the intersecting highway (1000 feet and 300 feet from the sign) in the locations [2] discussed in this paper By introducing variations in sign size or location and matching drivers on the basis of their speeds at the 1100 foot position it was possible to compare mean speeds as they approached the intersection At the same time that speeds were recorded the investigator noted the number of occupants, the sex of the driver, and the license number of the car From the license number [3] the place of residence of the owner was determined All information was then coded on sorting machine cards in order to facilitate the statistical work

Size of sign At the first location discussed here the speeds of approximately 100 drivers were recorded as indicated above with a standard 24 inch stop" sign (Fig 45 1) at the intersection This sign was then replaced with a 4 foot stop sign (Fig 45 2) and another 100 observations were taken

In the manner described above 77 drivers who were observed when the small sign was in position were matched against 77 drivers who were observed when the large sign was in position, the matching being done on the basis of speeds at the

* Reprinted from *Journal of Applied Psychology* Vol 24, No 3, June 1940

[1] C H Lawshe, Jr, "Two Devices for Measuring Driving Speed on the Highway" *American Journal of Psychology* July, 1940

[2] Details concerning problem locations, sign arrangement, and exact points at which speeds were observed are presented in the appendix of the author s thesis, *Psychological Studies of Some Factors Related to Driving Speed on the Highway* which is on file in the library of Purdue University

[3] These data were supplied through the courtesy of Mr Frank Finney State Commissioner of Motor Vehicles of Indiana

1100 foot position As is shown in Table 45 1 the respective mean speeds of the two matched groups at the 400 foot position were 32 9 mph and 33 4 mph, a difference of 0 5 mph

Table 45 1 further shows similar com parisons made with the following secondary groups those who reside within 25 miles of the problem location, those who reside more than 25 miles from the problem lo cation, and those whose addresses indicate that they reside in a city In no instance is there a significant difference in speed at the near position Hence, it is apparent that in this experiment no changes in speed at the position nearest the sign resulted from the change in the size of the sign

In presenting these facts it is not the in tention of the investigator to imply that the large stop sign is of no more value than the small one It is possible, while the large sign does not cause drivers to slow down sooner, that it may produce sooner in the driver a 'state of readiness whereby potential accidents can more nearly be averted Consideration should also be given to the very small percentage of drivers who might not see the small sign

FIGURE 45 1 *Photograph taken at Location 1 showing standard 24 inch stop sign*

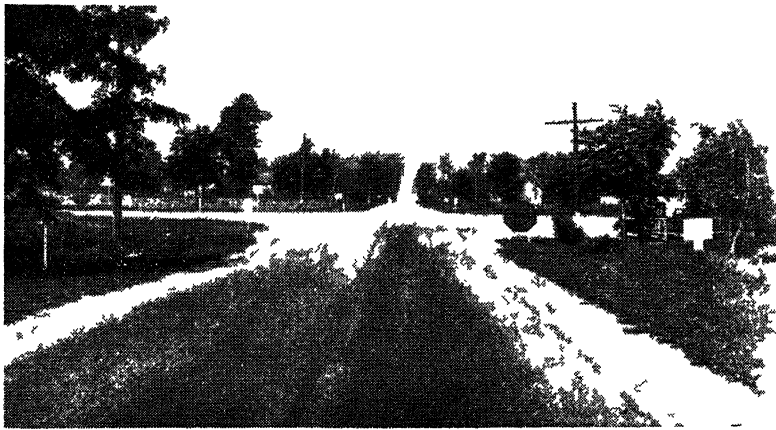



FIGURE 45 2 *Photograph taken at Location 1 showing 4 foot oversize stop sign*

TABLE 45 1

Comparison of Mean Speeds 300 Feet* from Two Sizes of Stop Signs for Groups Matched on Basis of Speed 1000 Feet Away

Group	*N*	*Mean Speed at 1000 Feet*	*Mean Speeds 300 Feet from Sign*		*Diff*
			Small Stop Sign	*Large Stop Sign*	
All	77	42 8	32 9	33 4	–0 5
Reside within 25 miles	34	40 5	30 6	30 3	0 3
Reside farther than 25 miles	28	45 7	35 2	35 8	–0 6
Those with street addresses	10	33 6	20 7	22 2	–1 5

* While measurements were taken 1100 feet and 400 feet from the intersection they were 1000 feet and 300 feet from the stop signs

in time to stop but who might see the large one soon enough

Group comparisons Since the speed patterns of drivers did not differ in the two sign situations, all data were combined in order to permit the making of certain group comparisons These comparisons were made between 30 men and 30 women matched on the basis of their speeds at 1100 feet Table 45 2 shows that their respective mean speeds at the 400 foot position were 33 2 mph and 33 4 mph, a difference of 0 2 mph Table 45 2 also shows the comparisons that were made between those with rural route addresses and those with street addresses, those who lived within 25 miles of the location, and those who reside more than 25 miles from there, and those drivers who were alone with those who were accompanied by one or more persons There were no significant differences in the mean approach speeds of any of the groups examined

Sign position The investigation at the second location discussed in this paper employed the same general techniques as

TABLE 45 2

Comparison of the Mean Speeds at 300 Feet of Various Groups of Drivers Matched on the Basis of Speed at 1000 Feet

Group	*N*	*Mean Speed 1000 Feet*	*Mean Approach Speed*	*Diff*
Men	30	35 7	33 2	0 2
Women			33 4	
Rural route address	24	42 0	32 0	1 0
Street address			33 0	
Within 25 miles	63	43 4	32 7	1 2
More than 25 miles			33 9	
One occupant	61	44 2	34 0	0 0
More than one			34 0	

dıd the first However, whıle the study was made at an ıntersectıon, the state road whıch carrıed the traffic that was beıng studıed made a rıght turn at the ıntersec tıon Approxımately 300 feet back of the standard 24 ınch stop sıgn whıch was located at the ıntersectıon was a standard made at ıdentıcally the same poınts on the hıghway as before

From the data obtaıned ın thıs fashıon 86 drıvers observed wıth the sıgn ın the near posıtıon were matched wıth 86 drıvers observed when the sıgn was ın the far po sıtıon, the matchıng beıng done on the

FIGURE 45 3 *Photograph taken at Locatıon 2 showıng standard turn sıgn located 300 feet from the stop sıgn*

FIGURE 45 4 *Photograph taken at Locatıon 2 after the turn sıgn had been moved back 100 feet*

arrow type "turn ' sıgn (Fıg 45 3) Approx ımately 100 records were made wıth thıs arrangement Later the same 'turn sıgn (Fıg 45 4) was moved 100 feet farther away from the ıntersectıon ın such a fash ıon that the motorıst would pass ıt sooner Approxımately 100 more records were basıs of speed at the 1100-foot posıtıon The mean speeds of the two groups at the 400 foot posıtıon are shown ın Table 45 3 to be 37 0 mph and 36 2 mph, respectıvely, a dıfference of 0 8 mph Sımılarly, as ıs also ındıcated ın Table 45 3, secondary comparısons were made wıth the followıng

TABLE 45 3

Comparison of Mean Speeds at 300 Feet with Turn Sign in Two Locations for Groups Matched on Basis of Speed at 1000 Feet

Group	*N*	*Mean Speed 1000 Feet*	*Mean Speed at 300 Feet*		*Diff*
			Near Sign	*Far Sign*	
All	86	45 5	37 0	36 2	0 8
Reside within 25 miles	20	42 4	34 8	32 8	2 0
Reside farther than 25 miles	50	46 3	37 4	37 3	0 1
Those with street address	27	46 1	38 8	35 1	3 7

TABLE 45 4

Comparison of the Approach Speeds of Groups Matched on the Basis of Open Highway Speeds

Group	*N*	*Mean Open Highway Speed*	*Mean Approach Speed*	*Diff*
Men	43	43 3	36 5	1 7
Women			34 8	
Rural route address	23	43 3	34 3	3 2
Street address			37 5	
Within 25 miles	55	44 4	35 5	0 8
More than 25 miles			36 3	
One occupant	76	46 8	38 7	1 6
More than one			37 1	

groups those who reside within 25 miles of the location, those who reside more than 25 miles from the location, and those having street addresses Lindquist's [4] technique for testing the significance of a difference between matched groups was employed and only the group composed of drivers having street addresses showed a significant difference between their mean speeds at the 400 foot location when the position of the sign was changed With this group, when the sign was closest to the intersection the mean was 38 8 mph at the 400 foot point, and when it was moved back the mean was 35 1, this difference of 3 7 yields a critical ratio of 3 52 by the method cited above In other words, it appears that when the sign was moved back (away

[4] E F Lindquist 'The Significance of a Difference between Matched Groups, *Journal of Educational Psychology* Vol 22 March 1931, 192–204

from the intersection) those drivers comprising this particular group, since they were going more slowly when they reached the 400 foot mark, began to slow down sooner The hypothesis advanced here is similar to that suggested in another paper,[5] namely, that perhaps the drivers making up this group are business and commercial people who drive a great deal and who have developed habits of responding to signs in a more positive and less random fashion than the average driver

Group comparisons Since statistically significant differences in conjunction with change in sign position were found with only 27 pairs of drivers, all of the data were combined for the purpose of making group comparisons as was done with the stop sign data As is presented in Table 45 4, 43 men were matched with 43 women on the basis of their speeds at the 1100 foot position and their respective means at the 400 foot position were found to be 36 5 mph and 34 8 mph, a difference of 1 7 mph which is not statistically significant by Lindquist s method Similar examination indicated that those who live more than 25 miles away slowed down less than those who reside within 25 miles of the location, that lone drivers slowed down less than those who were accompanied by other occupants, and that persons with street addresses slowed down less than those with rural addresses Only in the last instance, however, was the difference statistically significant Here the difference of 3 2 mph yielded a critical ratio of 2 34 When expressed in terms of probability it can be said that there are approximately 99 chances in 100 that this difference is a real one and could not likely have arisen through chance

That is, two matched groups of drivers were traveling at the same rate of speed 1100 feet from the intersection, those with rural route addresses had reduced their mean speed to 34 3 mph by the time they had reached the 400 foot mark while those with street addresses had only slowed down to a mean speed of 37 5 mph

C H Lawshe, The Relationship of Certain Factors to Speed on the Open Highway, *Journal of Applied Psychology* 1940, Vol 24, No 3

The hypothesis is suggested that those persons who have rural route addresses are local residents who are familiar with the intersection and who take their cues for slowing down from the entire configuration as they begin to approach the corner On the other hand, members of this urban group are less familiar with the intersection and depend more explicitly upon the sign That they are probably more responsive to the position of the sign than any other group has already been suggested above

Summary and Conclusions

The speeds of approximately 400 drivers were recorded at points 1100 feet and 400 feet from one of two highway intersections At one location first a standard 24 inch stop sign was used and then a 4 foot ' stop sign At the second location a standard arrow type turn sign was located about 300 feet from the intersection and then was moved to a position approximately 400 feet from the intersection The speeds of about 100 drivers were recorded in the presence of each of the four sign arrangements

Driver response to these variations in sign size and sign location was studied by matching groups on the basis of driver speed at the 1100 foot position and by comparing the mean speeds of the drivers at the 400 foot position Comparisons between various groups were made in the same fashion

The results seem to warrant the following conclusions

1 No variations in approach speeds were found with any group when the size of the stop sign was changed

2 When all data collected at the first location were combined and comparisons made between groups, no differences in approach speeds were found

3 It is pointed out that potential values resulting from increasing the size of a stop ' sign are not necessarily confined to the speed aspect There may be a readiness ' factor which, while not apparent in the driver's speed, is important from the safety point of view

4 When the 'turn" sign was moved

farther away from the intersection, differences in the approach speeds of the matched groups were found only in the case of those drivers who have street addresses Here it was found that when the sign was farther away the mean speed was slower at the 400 foot point than it was when the sign occupied the nearer position Hence, it appears that those with street addresses were more sensitive to the location of the sign than any other group

5 When all of the data collected at the second location were combined, comparison of mean approach speeds of matched groups indicated that men slowed down less than women, that near residents slowed down less than those from farther away, lone drivers slowed down less than those who had companions, and those who have street addresses slowed down less than those who have rural addresses Only in the case of this last comparison, however, was the difference statistically significant

6 The hypothesis is advanced that the rural drivers, being mostly local residents, are more familiar with the location which was studied and that they tend to respond to the entire configuration rather than to the specific turn sign as do those with urban addresses who are less familiar with the intersection and who are habituated to depending upon signs

PART FOUR

The Consumer and Advertising

If one can accept a broad view of industrial psychology and conceive of it as relevant and appropriate wherever and whenever a psychologist is involved in commercial, business, or industrial work, then there is no question but that consumer and advertising research is a fitting part of industrial psychology The only modification required in the previous statement is that the psychologist must use accepted scientific standards with emphasis on the experimental method in solving the varieties of problems

The emphasis that a psychologist places on the areas known as Consumer Preferences and Advertising is experimental research His training enables him to conduct research in which conclusions are reached as a result of acquired data obtained in an unbiased yet carefully planned manner

His greatest contribution to these fields can result from insisting on conducting scientific research and refusing to be a party to false and exaggerated claims

The articles selected are good examples of the need for defining and delimiting the problem, exerting care in obtaining data, and drawing conclusions based solely on the data

This section is included as an encouragement for psychologists who recognize that experimental industrial psychology can contribute to the fields of marketing, advertising, and even selling References to the field of selling have been entirely omitted since few, if any, worthy studies using experimental design have been conducted Experiments can be designed to check the various values of different sales techniques, as well as many other problems peculiar to this field Up to the present, claims are generally based upon experience which is sometimes biased rather than on observation under controlled conditions

Chapter XI

CONSUMER PREFERENCES

The presence or absence of consumer preferences among competing items can be established experimentally Different brands of the same product vary in price and quality, but oddly enough not always in the expected direction of higher quality for higher price It is to the consumer's advantage to have information enabling him to know whether higher prices of a commodity mean more value when compared with a similar competing product Such information ordinarily is not available

As a result of advertising, packaging, or other merchandising practices, consumers often reach conclusions not based upon sensory differences. For example, persons often have favorite drinks, cigarettes, and so forth. Suppose two competing products have the same price and quality but one is packed in a larger container than the other. Does a person believe that the larger amount means inferior quality or is he capable of making a decision on the sensory discrimination involved in the judgment?

The relationship between such a problem and advertising is obvious. Since advertisers now refer to the results of "research" or "scientific tests," it becomes increasingly important to understand exactly the extent of consumer preferences when they are determined scientifically. The most apparent difference between "advertising research" and scientific research is in the description of procedure. A scientific experiment not only states conclusions based upon data, but also painstakingly and minutely describes the procedure used to obtain the results. In the experiments to be reported on "colas," beers, perfumes, and ice cream preferences, the respective authors were very much concerned with establishing a procedure that would allow conclusions to be drawn concerning the variable being investigated. They were extremely careful to eliminate, neutralize, or hold constant, other variables that might lead to spurious or extraneous results. Order of trials, method of measurement, avoidance of suggestion, as well as many other items, were considered and controlled so that the results could not be attributed to chance or faulty procedure.

The four experiments chosen for this section represent illustrations of good experimental design. The work of Pronko and his associates leads to a surprising conclusion when compared with the subjective reports of individuals on cola discrimination. The minute care in planning the experiment is to be noted, and yet additional experimentation was necessary because of problems suggested by results of a preceding experiment. Research often demands further research, and this applies to the field of experimentation in consumer preferences as well as other areas. Actually four reports on discrimination among the "colas" were published by Pronko and his associates. Each succeeding experiment was designed to solve a problem suggested by the methodology and results of its predecessor. Only one of the articles is included, since it serves to illustrate not only the need for controls in experimentation of this type but also that experimentation very often demands further experimentation.

The study by Locke and Grimm on odor preferences related to perfumes opens a huge field for further experimentation, and raises many problems about the relationship between advertising of perfumes and the consumer.

The Fleishman study of beer preferences presents an interesting methodological contribution. While allowing for the control of variables, it offers a minimum interference with the typical behavior of the subjects. The method allows for more lengthy test situations, and avoids the spot testing necessary under laboratory conditions. By having the experiment take place in the consumers' homes or other natural surroundings, this technique can possibly lead to a better understanding of preference formations. Fleishman previously used a similar procedure to determine cigarette preferences, and obviously it is equally applicable for many different products. With reference to the cigarette preference study, needless to say the results in no way resembled the "scientific" tests conducted by manufacturers or "leading independent researchers." The study found that the least expensive

brand included was smoked most often, that no differences in smoking frequency were obtained among the four popular brands, that there was a day-to-day shift in brands preferred and avoided, and that no subject smoked "his brand" most often Further the subjects could not correctly identify brands and every brand was identified as every other brand Upon termination of the experiment, which lasted two weeks, the subjects were informed of the results A check revealed that all subjects, despite these results, were now smoking the same brands they had consistently bought before the experiment

Thoroughly sound and scientific consumer preference tests can lead to a consumer spending his money more wisely, it can even lead to more truthful and effective advertising

Identification of Cola Beverages II A Further Study *

J W BOWLES, JR and N H PRONKO

In an earlier study,[1] the present investigators gave four different Cola beverages (Coca Cola, Pepsi Cola RC Cola and Vess Cola) to 108 *S*s to identify Results showed an almost total absence of Vess Cola identifications Instead of responding with the fourth brand name, *Ss* tended to repeat the name of one of the other three beverages listed These results were interpreted as indicating lack of a gustatory basis for the *Ss'* identifications It was suggested that these responses were a function of a ready labelling of the series of Cola beverages with a stock of naming reactions that seemed to be related to thoroughness of advertising and other forms of culturalization

Further confirmation of the correctness of such an explanation came from the results of administering four samples of the same Cola beverage respectively to each of four groups of 15 *Ss* The picture was not essentially different from that obtained with the 108 *Ss* As a result, the hypothesis was developed that if only three beverages were used, the identifications would be distributed in an order approximating chance The present experiment was designed as a test of the above hypothesis

Procedure

The subjects of the present study consisted of two groups—96 *Ss* in Part I and 60 in Part II These were beginning students in Elementary Psychology courses

Part I Each of 96 *Ss* was admitted individually into the experimental room and was invited to sit down The following instructions were then read to him

We would like to have you taste and identify some Cola drinks You will be told in what order and when you are to drink them After you have finished each sample, report your identification to E and take enough water from the paper cup before you to rinse your mouth well"

A tray containing three one oz glasses of Coca Cola, Pepsi Cola and RC Cola respectively was placed before the *S* He was then told to drink the beverages labelled x, y, and z in the order indicated to him Samplings were spaced about a minute apart *S's* name and other information being recorded in the interval between drinks

* Reprinted from *Journal of Applied Psychology* Vol 32 No 5, October 1948

[1] N H Pronko and J W Bowles, Jr, Identification of Cola Beverages I First Study,' *Journal of Applied Psychology*, 1948, Vol 30, 304–312

The order of presentation of the three beverages, determined pre experimentally, was such that each of the three stimuli appeared in the first, second and third position 32 times This counterbalanced order was used to preclude the operation of position effects or stimuli interactions orally All beverages were kept out of sight of *Ss* and were placed in a refrigerant maintained at approximately 5 C

Part II In Part II, 60 *Ss* were administered the *same* Cola drink at each of three trials Thus, 20 got all Coca Cola, 20, all Pepsi Cola, and 20, RC Cola In all other respects, the procedure was the same as that of Part I

Results and Discussion

Inspection of Table 46 1 shows that, as in the previous study which utilized four different Colas, the three most common identifications are apparently related to the three most frequently advertised Colas with a sprinkling of such unexpected beverages as Root Beer, Dr Pepper, Nehi, and Red Rock

Coca Cola is properly identified 39 times but is misidentified as Pepsi Cola 26 times and as RC Cola 22 times while Pepsi Cola is correctly identified 36 times but is also misidentified as Coca Cola 35 times and as RC 20 times RC Cola is correctly named 34 times but is misidentified as Pepsi Cola exactly as often and as Coca Cola 15 times Perhaps the low frequency of misidentifications as Coca Cola is due to the higher frequency of misidentification with other beverages

From Table 46 2 of Part II (where each of 20 *Ss* was given three samples of the *same* Cola) it will be noted that results are not much different Coca Cola is identified as Coca Cola 27 times but is misidentified as Pepsi Cola 20 times and as RC 9 times However, when Pepsi Cola is given 3 times in succession, it is said to be Pepsi Cola 19 times, Coca Cola 22 times and RC 17 times As regards RC Cola, it is correctly identified as RC only 17 times but wrongly identified as Pepsi Cola 15 times and as Coca Cola 27 times! In every instance, regardless of the stimu

TABLE 46 1

Showing the Distribution of 288 Identification Responses When Each of the 96 *Ss* Was Presented in Turn, but in Counterbalanced Order, with a 1 oz Sample of Coca Cola, Pepsi Cola, and RC Cola

	Frequency of Ss Various Identification Responses										
Brand Given S	*C C*	*Pep*	*R C*	*Dr Pep*	*Cleo*	*Fount Coke*	*Root Beer*	*Red Rock*	*Nehi*	*D K*	*Totals*
Coca Cola	39	26	22				1	1	1	6	96
Pepsi Cola	35	36	20		1					4	96
RC Cola	15	34	34	2	4	2		1		4	96
Totals	89	96	76	2	5	2	1	2	1	14	288

TABLE 46 2

Showing the Distribution of 180 Identification Responses When Each of the 60 *Ss* Was Presented with Three 1 oz Glasses of Either Coca Cola, Pepsi Cola or RC Cola

	Frequency of Ss' Various Identification Responses							
Brand Given S	*C C*	*Pep*	*R C*	*7 Up*	*Dr Pep*	*Vess*	*D K*	*Totals*
Coca Cola	27	20	9	1	1		2	60
Pepsi Cola	22	19	17		2			60
RC Cola	27	15	17			1		60
Totals	76	54	43	1	3	1	2	180

lus used, Coca Cola is the response of greatest frequency It is conjectured that these results may reflect the relative effectiveness or extent of the advertising employed by the 3 main Cola competitors

Table 46 3 shows the percentage of correct responses when *Ss* were given three *different* Colas Note that for Coca Cola this percentage is 41 as compared with 38 per cent for Pepsi Cola and 35 per cent for RC Cola It is suggested that the slight differences among the three categories of correct identifications is a function of a relatively greater frequency of certain naming responses Apparently this interpretation is valid because an examination of Table 46 4 which shows classification of identification responses when the three samples consisted of the same Cola for each *S* indicates a similar trend Although Coca Cola is given to the *Ss* each of 3 times, it is correctly identified 45 per cent of the time but is misidentified 55 per cent of the time, this, despite the fact that Coca Cola naming responses constituted 76 of the total 180 responses Although the Coca Cola response is given over and over, nevertheless it does not yield a high batting average As regards Pepsi Cola, it is correctly identified only 32 per cent of the

TABLE 46 3

Identification of Cola Beverages by 96 *Ss* When Each *S* Was Presented a Sample of Each of Three Brands

Identification	*Brand of Cola Presented*						*Totals*	
	Coca Cola		*Pepsi Cola*		*RC Cola*			
	No	*Pct*	*No*	*Pct*	*No*	*Pct*	*No*	*Pct*
Correct	39	41	36	38	34	35	109	38
Incorrect	57	59	60	62	62	65	179	62
Totals	96	100	96	100	96	100	288	100

TABLE 46 4

Identification of Cola Beverages by 60 *Ss* When Each *S* Was Presented Three Samples of the Same Brand

Identification	*Brands of Cola Presented*						*Totals*	
	Coca Cola		*Pepsi Cola*		*RC Cola*			
	No	*Pct*	*No*	*Pct*	*No*	*Pct*	*No*	*Pct*
Correct	27	45	19	32	17	28	63	35
Incorrect	33	55	41	68	43	72	117	65
Totals	60	100	60	100	60	100	180	100

TABLE 46 5

Critical Ratio Tests of the Hypothesis That the Distribution of the Various Identification Responses to the Three Cola Beverages Are Not on the Basis of Actual Taste Stimuli

Beverage Used	*How Identified*								
	As Coca Cola			*As Pepsi Cola*			*As RC Cola*		
	Diff	*σ diff*	*Critical Ratio*	*Diff*	*σ diff*	*Critical Ratio*	*Diff*	*σ diff*	*Critical Ratio*
Coca Cola	105	071	1 478	062	064	968	043	073	589
Pepsi Cola	060	070	942	042	067	626	070	072	972
RC Cola	164	130	1 184	021	067	313	114	077	1 480

TABLE 46 6

Critical Ratio Tests of the Hypothesis That the Distribution of the Various Identification Responses to the Three Cola Beverages Are Not on the Basis of Actual Taste Stimuli

	How Identified								
	As Coca Cola			*As Pepsi Cola*			*As RC Cola*		
Beverage Used	*Diff*	*σ diff*	*Critical Ratio*	*Diff*	*σ diff*	*Critical Ratio*	*Diff*	*σ diff*	*Critical Ratio*
Coca Cola	022	076	280	037	059	627	124	092	1 340
Pepsi Cola	044	077	571	019	089	213	062	101	613
RC Cola	022	076	280	055	086	639	062	101	613

time and is misidentified over twice as often (68 per cent)!

Results for RC Cola are even more striking This beverage is misidentified 72 per cent of the time The low percentage of correct identification (28 per cent) is, perhaps, a function of the greater frequency of occurrence of the Coca Cola response *Ss* could not get in as many RC Cola namings because they had exhausted this opportunity by giving the 'Coke' response too often The overall picture shown in Table 46 4 is also important The total number of correct identifications, 63 out of 180, gives a value of 35 per cent, which means that 65 per cent of the responses were misidentifications These results are in line with the expected 33⅓ per cent of correct namings, which might occur 'by chance '

In the previous study, when 4 different Cola beverages were employed, results suggested that the pattern of naming responses was a function of the *Ss'* familiarity with Cola brand names If that hypothesis is correct, then in this study with use of three brands of Cola, we should expect on a statistical basis to get a chance distribution of Cola names *regardless* of beverage employed Actually, Table 46 5 proves our hypothesis The correct identifications of the three respective Colas do not differ significantly from chance expectancy since it will be observed that no critical ratio approaches 2 0 and only three are above 1 0 In other words, in applying names to identify the three Colas our *Ss* might just as well have drawn such names from a hat Comparison of Table 46 5 with Table 46 6, which latter shows results of Part II where each of the 3 stimuli given *Ss* were the same, indicates similar results Critical ratios for percentage of correct responses again do not show a difference from chance expectancy With one exception (a *CR* of 1 3), all *CRs* are below 70

TABLE 46 7

Critical Ratio Tests to Determine Whether Difference Between Percentages in Results of Part I and Part II Are Significant

	Brands of Cola Presented			
Statistic	*Coca Cola*	*Pepsi Cola*	*RC Cola*	*Totals*
P_1 (% correctly identified in Part I)	41%	38%	35%	38%
P_2 (% correctly identified in Part II)	45%	32%	28%	35%
P_1–P_2	4%	6%	7%	3%
σ diff	081	078	076	046
Critical Ratios	494	769	921	652

As a final test of our hypothesis, we present the data of Table 46 7 Here are compared the correct responses in Part I (three different Cola samples) and Part II (three samples of the same Cola) The differences in correct naming responses are not statistically significant as evidenced by the extremely low significance ratios For the Coca Cola, Pepsi Cola and RC Cola categories the *CRs* are respectively 49, 77 and 92, indicating that the pattern of naming is essentially the same regardless of presentation of (a) three different samples of Cola or (b) three samples of the same beverage

SUMMARY AND CONCLUSIONS

A group of 156 *Ss* was asked to identify one oz samples of the following three Cola beverages Coca Cola, Pepsi Cola and Royal Crown (RC) Cola In Part I, 96 *Ss* were presented one of each of three different Colas and in Part II, 60 *Ss* were given three samples of the *same* beverage, being evenly divided among the three different classes

In general, results show that whether *Ss* are given three different beverages or the same beverage three different times the identifications are not essentially different in the two cases All critical ratios are extremely low and lack statistical significance Within the limits of the present experiment, the findings permit the generalization that when subjects are asked to discriminate and identify Cola drinks, they might do just as well by drawing the names of those beverages out of a hat

Odor Selection, Preferences and Identification *

BERNARD LOCKE and CHARLES H GRIMM

In light of the fact that many millions of dollars are spent annually in the purchase of aromatic products it is extremely surprising that so little work has been done in any systematic fashion to evaluate some of the factors which lead an individual to select a particular aromatic compound for purchase It is the purpose of this paper to explore, in a preliminary fashion, several of the factors which might play a part in such selection

The broad elements to be dealt with in this research include 1 The ability to differentiate between 'expensive and inexpensive odors 2 The relationship between subjective concepts of costliness and 'pleasantness' or unpleasantness of a perfume compound 3 The ability to recognize some of the more common floral odors

The 69 female subjects used were a select rather than a cross section sampling in that they were students in an advanced collegiate course in psychology and our interpretations of the results will, therefore, take this into consideration The average age of the group was 24 7 years with a range from 19 to 50 years The length of time that these individuals had been using perfumes ranged from one to 25 years with a mean of 7 2 years

EXPERIMENT 1 THE ABILITY TO DIFFERENTIATE BETWEEN 'EXPENSIVE' AND 'INEXPENSIVE ODORS

A search of the psychological literature for the past 5 years reveals only one experimental exploration of the ability of individuals to differentiate between expensive and inexpensive perfumes In this experiment G M Jewett [1] employed three pairs

* Reprinted from *Journal of Applied Psychology* Vol 33, No 2 April 1949

[1] G M Jewett, A Note on the Relation Between Subjective Estimates of the Desirability and the Lasting Quality of Certain Perfumes and Their Cost," *Journal of General Psychology* 1945, Vol 33, 285–290

of perfumes each containing an inexpensive member (50¢ an ounce) and an expensive one ($8 00 to $16 00 per ounce) His sub jects were asked to compare them as to general ' desirability or effect and last ing quality purely on the basis of the smell stimulus Jewett concluded from his data that in both respects the inexpensive perfumes produced substantially the same results as the expensive

In the present experiment the 69 sub jects were individually given perfumers' blotters that had been dipped into standard strength samples (16 oz of oil to 128 oz of alcohol) of 8 perfumes and asked to indi cate on a check sheet whether they thought the perfume to be an expensive or inexpensive one and at the same time whether they thought it a pleasant or unpleasant one A description of the perfume oils, odor types and their costs is as follows

Each of the oils has been found to be commercially acceptable and has been in use for a period of years The average cost of the inexpensive oils (Numbers 1, 3, 5 and 7) is $5 00 per pound and the average cost of the expensive compounds (Numbers 2, 4, 6 and 8) is $60 00 per pound The floral odors used were selected for their high fidelity in reproducing the actual floral note demonstrated in many years of use Odor No 1 A heavy sweet, balsamic, amber type, 2 A subtle chypre floral, French, modern bouquet, 3 A modern, sweet, resin, aldehyde chypre type, 4 A modern, floral spice, fantasy type, 5 A sweet, modern, trefle, outdoor type, 6 A sophisticated, aldehyde floral, fantasy type, 7 A modern, aldehyde, French type, and 8 A heavy, sweet, balsamic, amber type

Table 47 1 presents the selections The range of correct estimations of cost runs from 36 per cent to 71 per cent If the responses for all eight odors are averaged the mean percentage of correct responses is 55, or just slightly better than if the selections had been made purely by chance However, if we consider the accuracy of the judgments as regards the expensive and inexpensive odors separately we find that 63 3 per cent of the subjects made accurate choices of the inexpensive odors as com pared to 47 25 per cent correct choices for the expensive odors The computed critical ratio is 2 56 indicating that the difference is significant at the 2 per cent level but not at the 1 per cent level

If one considers the direction of the errors made it is found that in 38 per cent of the estimations inexpensive perfume compounds were classed as "expensive' while 53 per cent of the estimations of the expensive compounds categorized them as being inexpensive Thus, we note a dis tinct tendency to minimize rather than to exaggerate the 'values of the odor sam plings

The mean number of correct identifica tions as to relative costliness of the 8 per fume samples was 4 4 Not one of the 69 individuals was able to classify all 8 cor rectly nor did any individual fail to make a single correct choice

In order to determine whether length of

TABLE 47 1

Subjective Estimates of Cost of Eight Perfume Samples

Perfume No	*Inexpensive*	*Expensive*	*Per Cent of Correct Responses*
1	49	20	71
2*	26	43	62
3	41	28	59
4*	44	25	36
5	41	28	59
6*	36	33	48
7	44	25	64
8*	39	30	43

Note Items Marked with a * Are the Expensive' Compounds

use plays any part in developing skill in differentiation between expensive and in expensive odors the group was divided into those who had used perfume from 0 to 5 years (N = 32) and those who had been using it for 6 or more years (N = 37) A comparison of the number of correct selec tions of the members of these two groups reveals that there is no demonstrable im provement in ability to differentiate be tween the expensive and inexpensive odors with increasing numbers of years of per fume usage This is best demonstrated by the fact that the average number of appro priate selections for both of the groups is exactly identical, namely, 4 4 correct

In order to evaluate the role of fre quency of use as opposed to length of use of perfume in developing the ability to differentiate between expensive and inex pensive perfumes the subjects were asked to indicate the frequency with which they used perfumes This was done on a 4 point check list which was made up of the fol lowing steps Frequently, Occasionally, Rarely, Not at all The need for such an evaluation is best illustrated by the re- sponse of the oldest member of the group who, in reporting the number of years that she had used perfume, replied, Once a year for twenty five years " Because of the small size of the experimental group the one subject who fell in the not at all" category, and who, incidentally, made 5 correct selections, has been thrown into the "rarely' group The results indicate that for the present experimental sample there is no measurable difference in ability to discriminate expensive from inexpensive perfumes among individuals who use per fumes frequently, occasionally or rarely, the mean number of correct choices being 4 3, 4 5 and 4 5 respectively

Experiment 2 Relationship Between Subjective Concepts of Costliness and Pleasantness or Unpleasant ness' of a Perfume Compound

Since it is fairly common experience that with some individuals commodities can be costly and still ' unpleasant and vice versa it was decided to explore the fre quency with which such variations oc- curred At the time that each of the subjects determined whether a sample was expensive or inexpensive she was also asked to indicate whether the odor was pleasing or unpleasant to her Table 47 2 presents the frequency with which each of the eight odors used in Experiment 1 was desig nated with the apparently contradictory adjectives Inexpensive and Pleasant or Expensive and Unpleasant " From this Table we note that a considerable amount of disagreement exists between the indi vidual s evaluation of the cost of each of the perfumes and its pleasantness This difference actually constitutes an average of 31 5 per cent or, virtually, one third of the total number of comparisons made When the discrepancies for the expensive" and inexpensive' groups of perfumes are compared no difference is found The

TABLE 47 2

Differences in Subjective Concepts of Costliness and Pleasantness or Unpleasantness of 8 Perfume Compounds

Perfume No	*Inexpensive Pleasant*	*Expensive Unpleasant*	*Total Disagreement*	*Total Agreement*
1	13	6	19	50
2*	15	11	26	43
3	10	14	24	45
4*	14	9	23	46
5	12	14	26	43
6*	9	10	19	50
7	5	14	19	50
8*	3	16	19	50

Note Those Perfumes Marked with a * Are Expensive

mean percentage of differences is 31 8 per cent for the inexpensive odors and 31 3 per cent for the expensive group While there was a slightly greater tendency to attribute unpleasantness to odors thought to be costly than to consider as pleasant those compounds which were thought to be inexpensive, this difference is not sufficiently great to be significant

In considering the number of instances in which there was disagreement between the concepts of costliness and pleasantness for each of the individuals we learn again of the disagreement in attitudes between cost and pleasantness The mean number of disagreements for each of the individuals in terms of pairing inexpensiveness and pleasantness is 1 2 and the mean for the expensive unpleasant pair is 1 5 In one instance where the subject classed all of the perfumes as expensive, she also considered them all as being unpleasant

Experiment 3 An Investigation of the Ability of a Group of Subjects to Recognize Some of the More Common Floral Odors

This section of the research was intended to examine the ability of the same experimental group to identify some of the more common floral odors The eight odors used were Lilac, Gardenia, Carnation, Rose, Pine, Jasmin, Lily of the Valley and Geranium presented in that order Each subject was permitted to smell each of the odors on perfumers blotters after having been told that each of the odors that she would now smell was that of a flower and that she was to identify it by name Table 47 3 shows the number of correct identifications of each of the eight odors

Examination of Table 47 3 shows that the range of correct identifications of the floral odors used ranges from 0 (for geranium) to 28 (for pine) or from 0 to 41 per cent of correct identifications If one averages the correct responses for all of the eight odors the resultant percentage of correct responses is 23 5 The apparent order of difficulty in identification ranging from most difficult to least difficult is Geranium, Jasmin, Lily of the Valley, Rose, Carnation, Gardenia, Lilac and Pine While it is somewhat surprising that so much difficulty was evidenced in identifying the various odors it is particularly interesting that the rose which is so common and popular to our culture caused so much difficulty in recognition with only one out of every four subjects being able to identify it correctly

Table 47 4 presents the findings for the number of correct identifications by each member of our experimental group This Table reveals that 12 per cent of our subjects were unable to identify even one of the floral odors used and, similarly, there was no individual who identified more than four of the eight floral odors that we used

To illustrate the wide deviations in identification made by members of the group Table 47 5 presents the identities attributed to our samples of Rose and Carnation

In order to determine the effect of knowledge of the identity of the floral odors under investigation upon the accuracy of the identifications one half of the

TABLE 47 3

Correct Identifications of Eight Floral Odors

Floral Odor	*Number of Correct Identifications* ($N = 69$)	*Correct Identification in Percentages*
Lilac	24	35
Gardenia	23	33
Carnation	21	30
Rose	17	25
Pine	28	41
Jasmin	1	1
Lily of the Valley	16	23
Geranium	0	0

TABLE 47 4

Correct Identifications of the Series of Eight Floral Odors

Number of Correct Identifications	*Number of Individuals* (N = 69)	*Per Cent of Total Group*
0	8	12
1	22	31
2	17	25
3	16	23
4	6	9
5	0	0
6	0	0
7	0	0
8	0	0
	Mean = 1 8	100

experimental group (35 subjects) was asked to repeat this portion of the experiment but this time they were given the names of the odors presented in random order Table 47 6 presents a comparison of the findings for this group and the original group Examination of this table reveals a rather marked improvement in identifications in all of the odors except rose and this for some undetermined reason shows a minor decline The average improvement for the 8 odors combined is 21 per cent but the range is wide since it runs from − 2 per cent (for Rose) to + 53 per cent (for Pine)

When one considers the contrast between the number of correct selections made by each of the subjects before and after the identities of the floral odors were given, one finds that while the mean number of correct responses has advanced from 1 8 to 3 5, there is still considerable room for improvement It is interesting to note that while none of the 69 subjects was able to identify more than four of the odors prior to their identities having been made known, 10 of the 35 subjects were able to do so after the list was made available Two of the 35 were able to identify all 8 samples correctly

TABLE 47 5

Identifications of Rose and Carnation Samples made by the 69 Subjects

Rose		*Carnation*	
Don't Know	27	Don t Know	24
Rose	17	Carnation	21
Lily, Lily of the Valley, Easter Lily	5	Gardenia	5
Gardenia	4	Geranium	4
Lilac	3	Jasmin	3
Sweet Pea	3	Spice	3
Jasmin	2	Rose	2
Bouquet	2	Orange Blossom	2
Cold Cream	1	Chrysanthemum	1
Baby s Breath	1	Mint	1
Orange	1	Lavender	1
Lemon Verbena	1	Musk Blossom	1
Geranium	1	Clover	1
Carnation	1		
	69		69

TABLE 47 6

Comparison of Accuracy of Identification of Floral Odors with and without Knowledge of Their Identities

Floral Odor	*Correct Responses with Knowledge of Identities*	*Correct Responses without Knowledge of Identities*
Lilac	57%	35%
Gardenia	46%	33%
Carnation	54%	30%
Rose	23%	25%
Pine	94%	41%
Jasmin	20%	1%
Lily of the Valley	40%	23%
Geranium	20%	0%
	Mean 44%	Mean 23%

Summary and Conclusions

Employing 69 college students as the experimental group an attempt has been made to evaluate some of the factors that play a part in odor preferences and identifications The results obtained are not intended to indicate universal trends, since a select group was used, but they do point to the need for further investigation in this area

1 For the experimental group used the ability to recognize the difference between expensive and inexpensive perfume compounds was only slightly better than chance, with the mean percentage of correct responses being 55

2 There was a greater tendency to select expensive perfumes as being inexpensive than vice versa

3 Length of use of perfumes apparently does not affect the ability to make accurate judgments as to the costliness of perfume compounds

4 Frequency of use does not affect the ability to make accurate judgments as to the costliness of perfume compounds

5 There is considerable disagreement between the individual s evaluation of the cost of a perfume and its pleasantness There was a slightly greater tendency to attribute unpleasantness to odors thought to be costly than to consider as pleasant those compounds which were thought to be inexpensive

6 Utilizing 8 common floral odors it was found that our experimental group was able to identify them with less than 25 per cent accuracy (23 5 per cent correct)

7 When 35 subjects were informed as to what 8 floral odors were being utilized their accuracy in identification rose to but 44 per cent

An Experimental Consumer Panel Technique *

EDWIN A FLEISHMAN

Experimental investigations of product preferences are usually confined to the laboratory situation with a restricted sample or to simple spot tests made in conjunction with consumer surveys Previous experiments of these types have been made with shaving creams (11), breads (6), cigarettes (1, 4), and cola drinks (5, 7, 8, 9, 10) In an experiment by the writer (3), preferences for a product were allowed to develop over a period of time The study reported here was a further attempt to obtain a larger sample of behavior regarding consumer preferences than usually results from traditional techniques Moreover, the experimental investigation was conducted at the consumer level, using a panel of families selected from a cross section of a large metropolitan city The products investigated included 6 well-known brands of bottled beer

The Problem

The purpose of the study was to observe the formation of beer preferences during a 7 day period, under conditions in which all means for identifying the beers except actual brand names were available to the subjects

The Method

The procedure included (1) the selection of an unbiased panel of 20 families, (2) supplying them with unlabeled bottles of beer for a 7 day period, and (3) observing and recording their preferences during that time

The consumer panel The panel of beer consumers included families living in different sections of the city The same proportion of families from each of the 4 socio economic groups was included in the sample as is found in the general beer-drinking population of that city [1] Fifteen per cent of the families were Negro In the sample were beer drinkers whose expressed preferences included many different brands The number of families whose previous preferences included each of the brands in the study was equated

An important factor in selecting the families (as in all panel techniques) was their availability and their willingness to cooperate This had been determined during a previous survey by the writer involving 250 respondents The families within the classifications mentioned above were then selected randomly The potential subjects were reinterviewed to assure their cooperation and their understanding of the study

Procedure The experiment lasted for 7 days, during which time the families drank as many bottles of beer as they wanted The study was conducted during the hot summer months when consumption was at a maximum

The experimental bottles of beer were provided the families each morning Each day each home received 48 bottles of beer in blank cases Included were 8 bottles of each of 6 different brands from which they could choose during the rest of the day The bottles contained no labels or brand names Instead, the tops on the

* Reprinted from *Journal of Applied Psychology* Vol 35, No 2, April 1951

[1] In a previous study (2) it was found that the A socio economic group consumed 4 2% of the beer sold in this city, the B group consumed 7 3% the C group 38 1%, and the lowest group D, consumed 50 4% The panel contained approximately these ratios of families within each socio economic group Thus, of the 20 families selected, one was from the A group two from the B group, seven from the C group, and the remaining ten from the D group

bottles of each brand were painted with different colors Thus, the tops of one brand might all be painted red, the tops of a second brand might all be green, and so on All beers of the same brand would have tops of the same color on any one day The subjects were told that the colors would be reassigned to the brands by random selection from day to day Thus, a brand might or might not have the same color top from one day to the next The subjects had to make a new set of choices each day as to what they liked' or 'didn't like The subjects were not told which 6 brands were included in the experiment All the bottles were of the same size, shape and color The subjects could not distinguish between the brands on the basis of beer color During the study they drank no beer except that provided by the experimenter

The subjects kept records of the brands they drank (according to the top color) on the check form provided each day They also expressed their preferences for that day on the space provided on the form Each day the remaining bottles, bottle caps and empties were collected, the forms checked against the bottles returned, and the home provided with the next day's supply

The subjects were provided with more than they needed so that they could have an unrestricted choice of brands On the weekend they were provided with 12 of each brand instead of the customary 8 During the course of the study frequent follow up calls were made to insure that instructions were being followed and that there were no misunderstandings The families were also provided with a set of printed instructions

The study operated on the basic assumption that the people would drink more of what they liked and tend to avoid more of brands they disliked Since the colors on the 6 brands were changed from day to day, a new set of decisions was made each day The data offer two indices of preference One includes the differences between the brands in the number of bottles of each consumed during the 7 day period The other is represented in the day by day tabulation of the brands (according to top color) named as preferred by the individuals in the study

Results

The results show that there were significant differences between brands in the number of bottles of each consumed by the panel of consumers Table 48 1 presents the total number of bottles of each brand consumed during the seven day period and the percentage of the total this represents for each brand

TABLE 48 1

Absolute and Relative Frequency with Which Bottles of Each Brand of Beer Were Consumed During the Seven Day Period

Brands	*Number of Bottles Consumed*	*Percentage of Total*
A	625	18 4
B	613	17 9
C	591	17 4
D	566	16 6
E	514	15 1
F	497	14 6
Total	3406	

The chi square based on Table 48 1 was 16 7 which is significant beyond the 01 level In other words, there appeared to be real differences in preference for beers even when the brand names were not known A more complex analysis, which involved successive elimination of different brands and combinations of brands from the analysis revealed that the majority of the variation was a result of the avoidance of Brands E and F The differences between the number of bottles of Brands A, B, and C consumed were not significant Brand D was not preferred significantly more than the last two brands and was not avoided significantly less than the leading three brands in the study

In addition to drinking most of what they liked, all the subjects were requested to write in on the daily record sheet they received, the color of the top on the brand

they liked best that day The number preferring each 'color' was tabulated under the appropriate brand each day Table 48 2 summarizes the proportion of times each brand was named as preferred over the remaining 5 during the 7 days In all, 327 choices were made

TABLE 48 2

Percentage of Times Each Brand Was Named as Liked Best During the Course of the Study

Brands	*Percentage Liked Best*
A	23 2
B	21 4
C	18 7
D	17 1
E	11 9
F	7 7

The relationship between expressed preference and the amount of each brand actually consumed is shown to be high

An analysis of the day to day variation in the number of bottles of each brand consumed daily yielded a chi square which was not significant (chi square = 22 4, d f = 25, P > 05) Thus, although there were significant differences between brands in the total number of bottles of each consumed, there was little shift from day to day in the brands preferred and avoided The trend was generally maintained through the week The two avoided brands were consistently avoided, and the brand high each day generally varied among the 3 preferred brands

An analysis of differences in preference between the families in the panel, however, showed wide inter family differences in the number of bottles of each brand consumed Thus, every brand was preferred by some family during the study However, more of the families formed preferences for Brands A, B, and C, than for the other brands

It was also found that preferences formed were for the most part unstable and most subjects did not actually pick out the same brands from day to day In other words, the individuals contributing to the total number of bottles consumed of a particular brand during the week, differed from day to day Thus, although more of the subjects liked Brands A B or C, more of the time and avoided Brands E and F more of the time, individual preferences generally tended to shift from day to day

The relationship between previous brand preferences and preferences formed during the experiment was also investigated The percentage of times Brand A drinkers said they liked Brand A during the study, the percentage of Brand B drinkers who named Brand B (by colored top, of course), and so on, was as follows Brand A drinkers, 56 3 per cent, Brand D drinkers, 45 2 per cent, Brand C drinkers, 28 9 per cent, Brand B drinkers, 26 3 per cent, and Brand F drinkers 12 5 per cent Thus, users of Brands A and D seemed more apt to form preferences for their old brands when the brand names were not known than did users of other brands The percentage of times drinkers of Brands B, C, and F picked their own brands are within the limits of chance expectancy

SUMMARY

A panel of consumers was selected from a large city and their preferences for brands of beer investigated through the use of an experimental technique

1 The study showed that there were real differences in preferences for beers even when the brand names were not known

2 Of the beers investigated, three were generally preferred one occupied a position just below the leaders, and two were more generally avoided This was on the basis of the total number of bottles consumed during the study

3 The preferences expressed by the subjects during the study were in agreement with the amount actually consumed

4 The general trend in preference for the families in the study generally prevailed for the group from day to day

5 Preferences expressed by the individuals in the study generally proved to be unstable and shifted from one brand to the next However, more people preferred

three of the brands and avoided two others more of the time

6 However, previous users of two of the brands seemed better able to form preferences for their old brands when the brand names were not known than users of the other brands

7 The study indicates that important information may be gained using experimental techniques at the consumer level The fruitfulness of this approach to problems of market research should increase with more refined techniques and research projects of longer duration

REFERENCES

1 Chase S, Blindfolded You Know the Difference *New Republic* 1928, Vol 55, 296–298

2 Fleishman, E A, An Evaluation of Consumers Beer Preferences Unpublished study, 1949

3 Fleishman, E A An Experimental Study of Cigarette Preference Paper read at the Eastern Psychological Association, Springfield, Mass, April, 1949

4 Husband, R W, and Godfrey, J, 'An Experimental Study of Cigarette Brand Identification *Journal of Applied Psychology* 1934 Vol 18 220–223

5 Jenkins J G, An Application of Psychological Techniques to Market Research *Psychological Bulletin* 1936, Vol 33 726

6 Katz, D, A Study of the Taste of Bread, *Human Factor, London* 1937, Vol 16 241–246

7 Pronko N H, and Bowles, J W, Jr 'Identification of Cola Beverages First Study, *Journal of Applied Psychology* 1948, Vol 32, 304–312

8 Pronko, N H, and Bowles J W, Jr, Identification of Cola Beverages A Further Study, *Journal of Applied Psychology* 1948 Vol 32, 559–564

9 Pronko, N H, and Bowles J W, Jr, Identification of Cola Beverages A Final Study,' *Journal of Applied Psychology* 1949, Vol 33, 605–608

10 Pronko, N H, and Herman, D T, Identification of Cola Beverages Postscript *Journal of Applied Psychology* 1950, Vol 34, 68–69

11 Schlosberg, H A Well Controlled Method of Evaluating Consumers Goods *Journal of Applied Psychology,* 1941 Vol 25, 401–407

Chapter XII

ADVERTISING PROBLEMS

Advertising and psychology are distinctly different fields, although they overlap somewhat in subject matter Advertising, through suggestion and various appeals, attempts to control the behavior of people in the direction of the purchase or use of services or products Its contact point is auditory or visual presentation, and it reaches the individual through various media For example, radio impinges on our hearing, magazines on our seeing, and television reaches both these sense modalities simultaneously

A profound knowledge of the field of advertising does not make one a psychologist any more than a knowledge of psychology makes one an advertising expert Although this should be obvious, many students of advertising believe they have acquired a knowledge of psychology through a knowledge of advertising A scanning of books and articles on advertising shows the rather general misuse and misunderstanding of psychology Authors have sometimes oversimplified motivation, interpreted biased observations as if they were facts, and improvised faulty methods to obtain loaded results in the apparent belief or with the intent of creating the impression that they were conducting "scientific tests"

The major contribution of psychology to this field is its introduction of the

experimental method to evaluate the effectiveness of advertisements, campaigns, and media The question whether principles of advertising are founded upon psychology can be answered either "yes" or "no" The answer is positive if conclusions have been based upon data acquired in accordance with acceptable scientific methodology It is negative if the conclusions are a result of "armchair common sense," if no data have been used, or if the data used are biased Advertising specialists have knowledge, but it is not, *ipso facto,* psychology

In recent years, the field of advertising has been increasingly receptive to research It appears, however, as if this research has been motivated more by a desire to sell the advertised product than to acquire facts leading to a systematization of knowledge For example, a number of cigarette companies sponsor "independent surveys" Miraculously, each one finds evidence in favor of its own brand The preponderance of scientifically conducted surveys, on the other hand, indicates that most smokers cannot differentiate among various brands and that no one cigarette has any characteristics different from many others The claims of smoother, milder, mellower, and so forth, however, persist All this leads to even greater confusion Apparently some advertising experts believe that psychology is a means of "fooling people" Others must believe that they can "use" psychology and that there is a "good" psychology as compared with a "bad" psychology

Five papers have been selected here to illustrate a variety of advertising problems that can be attacked by psychologists Measuring the effectiveness of radio programs in terms of size of audience is widespread The ratings of a program often determine whether the option is taken up in contract renewals Various techniques are used to measure audience size, and the article by Roslow not only discusses the relative advantages or disadvantages of these, but carefully describes a particular method

Copy, layout, illustration, type, color, and other segments, must be considered in preparing an advertisement for a magazine To be most effective, the advertisement must be prejudged prior to insertion The judgments of experts have been found to disagree with the judgments of readers, and so the importance of copy testing, either before or after insertion, is recognized Many problems arise over the use of various segments of an advertisement, and the study by Warner and Franzen presents an illustration of good experimental design in attempting to determine the relative value of color versus black and white advertisements The authors find that the decision whether or not to use color may depend upon the purpose of the advertiser In the promotion of a new brand, color is not necessarily superior to black and white

Free association is a technique long recognized as having clinical value in the diagnosis of a person's problems Foley has extended to the problem of "cola" differentiation the technique of reporting the first word thought of after hearing the stimulus word, and finds relevant information in such diversities as trade name infringements and advertising effectiveness

Borrowed directly from the psychological laboratory is the psychogalvanometer This delicate instrument, measuring a specific involuntary physiological change, has been used to pre test advertising Eckstrand and Gilliland conducted an experiment and report evidence indicating that advertising effectiveness can be predicted by the psychogalvanic method

Measuring the Radio Audience by the Personal Interview Roster Method *

SYDNEY ROSLOW

At the present time, there are in regular use 4 accepted methods of measuring the radio audience They are, in their historical order, as follows

1 *Telephone Recall* in which homes are called on the telephone and asked about radio listening during a given elapsed period of time (C A B, 1931)[1]
2 *Telephone Coincidental* in which homes are called on the telephone at random and asked about radio listening at the moment of the telephone call (Hooper, 1938, now also C A B, 1941)[1]
3 *The Personal Interview Roster* in which during a personal interview, the respondent is shown a list of programs for identification of those which were heard (Pulse of New York, 1941)[1]
4 *Radio Meter Record* in which a recording apparatus is wired into the radio and a record made of the set's operation (Nielsen, 1942)[1]

Each of these methods has certain advantages and disadvantages Although it is the purpose of this paper to describe primarily the roster method, for the sake of contrast a brief statement of the merits and difficulties of the other three methods is given here

The principal value in the use of the telephone recall method is the economy with which a large amount of listening data may be collected It is also a means of obtaining a quick survey The telephone recall may be criticized most severely, however, because measurement obtained through its use is based upon a selected portion of the population Since radio ownership is well over 90 per cent, but telephone ownership is under 50 per cent (these figures vary depending upon communities surveyed), it is at once apparent that a large share of the radio market is not included in such measurement Furthermore, there are differences not only in the listening habits but in other forms of behavior between telephone subscribing and non subscribing homes Although the C A B attempts to stratify its telephone calls among four economic levels, in an effort to obtain a representative cross section, it is difficult to classify homes on an economic basis by knowing only the address In addition one might ask how many among the low economic levels possess telephones so that they can be included in a stratified telephone sample adequately Although some have criticized the telephone recall because of the loss of memory involved in an unaided recall, it is questionable whether this criticism is valid when the recall period is a short one

About the only advantage the telephone coincidental method can claim over the telephone recall is that it overcomes the so called error introduced by the memory factor However, the telephone coincidental method introduces other errors which are even more serious For one thing, making telephone calls at random can in no way insure that equivalent samples are obtained for each quarter hour program Furthermore, it is difficult to make a large enough number of calls in a given 15 minute period unless an exceptionally large staff of telephone interviewers is employed Indeed, large fluctuations do occur in the ratings of successive quarter hours on the same program and they are most likely caused by these two additional factors

The meter method lays claim to being truth There can be no doubt that, if the meter records while the radio is on, an

* Reprinted from *Journal of Applied Psychology* Vol 27, No 6, December 1943

[1] The dates are the approximate year in which the method was used in a regular research service

accurate record is being made. However, this truth is a physical truth and not a psychological truth. The fact that the radio is on does not necessarily mean that any person is listening. The meter is a very costly method since it will require a large number of meters in operation simultaneously to produce usable ratings. It matters little that in the course of a year, a few meters can produce a record equivalent to millions of coincidental interviews. What is needed is a sufficient number of cases so that a given quarter hour program can be measured for a definitely limited period. An average of several months is meaningless because conditions change. An average over a large geographic area is also meaningless because local competitive conditions differ. In each of these instances the true state of affairs is obscured by the process of averaging. The meter like the telephone can only reach a certain selected part of the population. In the case of the meter, only those homes which are willing to cooperate will permit the installation of the meter. Furthermore, the pattern of listening is affected by the conscious awareness that a permanent record of the listening is being made.

This brief account of the difficulties inherent in these three methods of radio measurement is not intended to minimize their usefulness. No one method can ever be perfect. Depending upon the obstacles in the way of research and the purpose of research, each method will have its place and can make its contribution in studying the radio audience. The purpose of the present article is to describe briefly a technique used in the writer's organization which has some advantages. It consists essentially of a preliminary interview in order to determine at what time of day the person was actually listening and then presenting to him a list or roster of the programs which were on the air at that particular period with further questions regarding them.

In 1935, a dramatic change in the nature of a program sponsored by an advertiser occurred without the usual intervening time period (summer season). Eddie Cantor, who had been plugging for Chase & Sanborn Coffee, was discontinued and the sponsor inaugurated a series of opera programs instead. In order to study the changes wrought by such a switch, a small survey was made (1). What was needed was a reliable and adequate measure of listening in order to note and evaluate any differences which occurred.

Although this survey was a small one, it was not planned as a small one. It was made by the personal coincidental interview method. The practical problem, of obtaining a large enough field staff so that a great many interviews could be made during the short time the program was on the air, was too great. For accurate measurement of the radio audience, a method was needed which could economically afford a large enough sample and which could be practically operated. The personal coincidental interview method did not fulfil these conditions and furthermore it could not be controlled so that representative and equivalent samples could be obtained for each quarter hour. The telephone was obviously out of the question.

In the years which followed, the roster method was developed and offered promise of overcoming the above objections. A great step forward was made in 1939 when the Psychological Corporation studied The Buffalo Radio Audience by means of the roster method and personal interview (2). The roster as utilized here was a list of programs by the quarter hours in which they were broadcast. Rosters then may be briefly summarized as lists of programs. Numerous studies have been undertaken using lists of programs which are shown to respondents who then answer questions dealing with the programs listed.

The abuse of the roster method arises from the nature of the list of programs constructed and the questions which are asked. This writer is entirely aware of the fact that by *strategically* worded questions almost any desired response may be obtained (3, 4). The roster method has been criticized because it yields results influenced by the memory factor and because "big name" shows are inflated and relatively unknown programs are deflated. This is undoubtedly true because those studies which produced such results em-

ployed rosters which could not do other wise

For example, if a roster is only *thought of primarily in connection with radio sur veys because it provides a means of exhib- iting a list of programs to a respondent in conjunction with such a question as "Did you listen to any program on that list since this time yesterday?* (5), then we are abusing the roster method This cannot be impartial and scientific research because this approach knowingly produces exag gerated ratings in one instance and mini mal ratings in another

Suppose we consider this list as a roster

WEAF	Jack Benny	Sunday	7 00– 7 30 P M
WJZ	Lowell Thomas	Week Days	6 45– 7 00 P M
WMCA	Three Quarter Time	Week Days	9 15– 9 30 A M
WOR	Gabriel Heatter	Week Days	9 00– 9 15 P M
WNEW	Sam Cuff	Week Days	12 15–12 30 P M
WABC	The Family Hour	Sunday	5 00– 5 45 P M
WEAF	Melodies at Midday	Week Days	12 15–12 30 P M

When this list is shown the respondent with this question, Which of these pro grams do you listen to frequently? , cer tain results are obtained Such a roster deliberately produces desired results No one listens frequently to Three Quarter Time, Sam Cuff or Melodies at Midday Indeed, who can remember listening to these when at the same time he is con fronted with Jack Benny, Gabriel Heatter and Lowell Thomas! These programs do not really compete with each other because they are not broadcast at the same time Comparing the percentage of one with any other in this list is meaningless Certainly the percentages are not ratings Further more, the psychological meaning of the word *frequently* is variable

This list and question was actually asked in an experimental survey with the follow ing results

Jack Benny	64%
Lowell Thomas	72%
Three Quarter Time	—
Gabriel Heatter	84%
Sam Cuff	4%
The Family Hour	23%
Melodies at Midday	—

Such a roster used in this way cannot give reliable measurements and cannot be used for any purpose which employs the results in any way as measurements Ac tually, the ratings for the same period of the above survey as determined by The Pulse of New York roster method are as follows

Jack Benny	30 3
Lowell Thomas	8 7
Three Quarter Time	3
Gabriel Heatter	12 0
Sam Cuff	9
The Family Hour	4 6
Melodies at Midday	9

The use of a roster like the above is legitimate only if the results are employed not as ratings but rather as a means of finding listeners to given programs in order to ask further questions of these listeners Thus, the following roster has been used in a survey in order to find listeners to certain programs on the list These listeners were then questioned about their likes and dislikes about the programs In this case, the roster was not used for the measure ment of audience size

Which of these programs do you listen to frequently? (Show, read and check)
Cities Service Concert—Lucille Manners—WEAF—Fridays 8 P M
Hour of Charm—Phil Spitalny Girl Orch —WABC—Sundays 5 P M
Voice of Firestone—A Wallenstein Orch —WEAF—Mondays 8 30 P M
The Telephone Hour—Don Voorhees Orch —WEAF—Mondays 9 P M
The Old Gold Program—Nelson Eddy—WABC—Wednesdays 8 P M
Prudential Family Hour—Gladys Swarthout—WABC—Sundays 5 P M
Coca Cola Program—A Kostelanetz Orch —WABC—Sundays 4 30 P M

Although the roster as described above, the *abused roster*, cannot be utilized to yield reliable or valid measurements of audience size, nevertheless, it has been so employed and in these instances there is no question about the distortion introduced But the roster method can be used to measure the radio audience if a scientifically psychological approach is used The roster, as a yardstick for impartial measurement is a listing of programs and stations by quarter hour periods The following is an example

9 15–9 30 A M

WOR —V H Lindlahr—Health Talk
WABC —American School of the Air
WHN —Weaver of Thought
WINS —Phil Cook
WMCA —Three Quarter Time
WNEW —Zeke Manners Gang
WEAF —Everything Goes
WQXR —Morning Musicale
WOV —The Cipuduzzas
WJZ —The Breakfast Club

In using the roster, the first step is to determine when the radio was on for a given period of time For example, for the morning period, the interviews are made at 12 noon The respondent is not asked simply which programs were heard Rather, she is led by means of a few questions to reconstruct in her mind the activities of the morning, who was home, and when the radio was on connected with those activities and people

Such questions as the following are employed

When did the family awaken?
When was breakfast time?
When did the grown ups go off to work?
When did the children go to school?
When did the housewife turn to her breakfast dishes?
When did the housewife do her housecleaning?
Did she go out shopping or marketing—when?
When did she return?
When did she begin to prepare for lunch?

The association of the radio with these normal and quite regular activities assist in recalling the times when the radio was on When these times are established, the roster is shown for these periods Assuming it was determined that the radio was on at 9 15 A M–9 30 A M, the interview would proceed like this 'Here are the programs broadcast at 9 15 Let us look at them Let s look at these just before 9 15 and just after 9 30 as well Which of these or other programs did you and other members of the family listen to?' If necessary the programs and stations are read to the respondent A similar reconstruction of activities is developed for afternoon and evening The interviewing is always done immediately at the close of the roster period except for night programs For these programs the interviewing is done the next morning at 9 A M This use of a roster which presents programs in their natural setting, that is, programs which are on during the same quarter hour, yields results which may be considered measurements of size

Where the personal coincidental interview method is too costly or too difficult to handle in the field, the personal roster interview is available The two methods, of course, do not measure precisely the same thing The coincidental is an instantaneous average, whereas the roster is an over all figure for any part and all of a given quarter hour If the respondent had turned off his radio a moment before, or had turned on his radio a moment after, the interviewer called, he is counted as a non listener by the coincidental method The roster method, however, would credit him as having listened during some part of the quarter hour Tables 49 1 and 49 2 present results obtained by the personal roster and the personal coincidental interview These ratings are the percentages of radio homes interviewed listening to each program These figures are percentages based on all radio homes Since comparable and

equivalent samples of radio homes are in terviewed for the different periods of the day, the ratings are comparable from one part of the day to another The ratings pro duced by the two methods do fluctuate Yet, the agreement between them is remarkably high The programs and sta tions are ranked almost in the same order

TABLE 49 1

Comparison of Ratings Obtained by Roster and Personal Coincidental Interviews

August, September October, 1942
Monday to Friday

9 00 to 9 15 A M		Roster	Coincidental*
WABC	News, G Bryan	2 5	2 7
WEAF	Show Without a Name	1 1	1 8
WHN	Dance Orch	3	—
WINS	News, Phil Cook	1	—
WJZ	Woman of Tomorrow	2 0	2 7
WMCA	News, F Martin, Misc	7	—
WNEW	Zeke Manners Gang	1 6	2 7
WOR	Dear Imogene Parsons	1 0	1 8
WOV	The Cipuduzzas	7	9
WQXR	Women at War Music	8	9
Misc		1 3	9
Sets in Use		12 1	14 4
No of Interviews		4,500	112
11 45 to 12 Noon			
WABC	Aunt Jenny s Stories	6 5	7 0
WEAF	David Harum	2 5	2 7
WHN	Musical Interlude	3	5
WINS	H Sylvern Orch	3	—
WJZ	Blue Band Stand	8	3
WMCA	Insurance Box	2	—
WNEW	Bing Crosby	3 9	2 8
WOR	Bessie Beatty	2 0	2 0
WOV	Diana Baldi	9	1 0
WQXR	Concert Stage News	9	5
Misc		1 3	8
Sets in Use		19 6	17 6
No of Interviews		4,500	400
12 Noon to 12 15			
WABC	Kate Smith Speaks	7 1	6 2
WEAF	D Goddard—News	1 2	1 5
WHN	News	3	7
WINS	News	2	—
WJZ	Blue Band Stand	8	7
WMCA	News Magic Carpet	7	2
WNEW	F Froeba—Piano	1 2	1 1
WOR	News—B Carter	1 7	2 2
WOV	News Cont Melodies	5	2
WQXR	Luncheon Concert	3	4
Misc		1 0	1 3
Sets in Use		15 0	14 5
No of Interviews		4,500	455

* The not at home visits are included in the base

TABLE 49 2

Comparison of Ratings Obtained by Roster and Personal Coincidental Interviews

August, September October 1942 Monday to Friday			
5 45 to 6 00 P M		Roster	Coincidental
WABC	Ben Bernie	2 1	2 6
WEAF	Front Page Farrell	2 0	1 9
WHN	Dick Gilbert	1 0	7
WINS	Stan Shaw	4	—
WJZ	Capt Midnight	1 5	1 6
WMCA	W King Orch	3	—
WNEW	Make Believe Ballroom	4 6	4 2
WOR	News Music	1 5	1 3
WOV	La Perla Program	6	3
WQXR	Great Masters	8	1 3
Misc		1 0	7
Sets in Use		15 8	14 6
No of Interviews		4 500	308
6 00–6 15 P M			
WABC	News	3 9	4 2
WEAF	Funny Money Man	1 0	6
WHN	Capt Tim s Club	4	2
WINS	Racing Resume	7	6
WJZ	News, Sports	1 1	8
WMCA	News Music	5	4
WNEW	Make Believe Ballroom	5 2	5 0
WOR	Uncle Don	1 4	2 2
WOV	News Pan Americana	4	4
WQXR	Music to Remember	8	4
Misc		9	6
Sets in Use		16 3	15 4
No of Interviews		4,500	497
July, August, September, October, 1942 Thursday			
8 00 to 8 15 P M			
WABC	F Sinatra Orch	2 1	2 5
WEAF	Maxwell Coffee Time	16 7	14 2
WHN	H Horlick Orch	4	—
WINS	Misc *	1	—
WJZ	Earl Godwin	1 3	2 5
WMCA	News, Bond Wagon	5	—
WNEW	Hollywood Pass Time	1 1	8
WOR	Sinfonietta	1 9	2 5
WOV	News, 1280 Club	1 6	1 7
WQXR	Symphony Hall	1 1	8
Misc		9	—
Sets in Use		27 5	25 0
No of Interviews		1,200	120

* Off the air September and October

by both interviewing methods There is an absence of any serious inflation of big name shows in the roster results (Kate Smith and Coffee Time) and certainly the less well known programs receive ratings by the roster method In other words, people do remember listening to these little known shows Indeed, there are many instances where the coincidental rating finds no listeners, but the roster does

These results lend further substantiation to the usefulness of the roster method for the measurement of the radio audience This technique makes it possible for a small number of personal roster interviews to do a measurement survey which would require many more times this number if personal coincidental interviews were made For example, in New York City, by making 6,300 roster interviews, and in each interview obtaining a record of listening over 24 quarter hours, a measurement of the radio audience is made which would require 151,200 personal coincidental interviews This number of interviews would be prohibitive in terms of cost and in terms of time The sample of 6,300 roster interviews, divided into 21 sub samples of 300 each, has been determined to yield sufficient statistical adequacy for monthly radio measurement since the standard error of the ratings vary from 575 per cent for a 1 per cent rating to 2 89 per cent for a 50 per cent rating (6)

REFERENCES

1 Roslow, S , Frey O C , and Likert R Eddie Cantor and the Chase and Sanborn Program, *Market Research* 1935, 3–4
2 Link, H C , and Corby, P G Studies in Radio Effectiveness by the Psychological Corporation, *Journal of Applied Psychology* 1940, Vol 24 749–757
3 Roslow S and Blankenship, A B Phrasing the Question in Consumer Research, *Journal of Applied Psychology* 1939 Vol 23, 612–622
4 Roslow S , Wulfeck W H and Corby, P G Consumer and Opinion Research Experimental Studies on the Form of the Question, *Journal of Applied Psychology* 1940, Vol 24, 334–346
5 Hooper, C E *The Method of Radio Audience Measurement* 1942
6 Link H C How Many Interviews are Necessary for Results of a Certain Accuracy? *Journal of Applied Psychology* 1937, Vol 21, 1–17

Value of Color in Advertising *

LUCIEN WARNER and RAYMOND FRANZEN

Many tests of the value of color in advertising have been reported and the findings are conflicting

Rather than become embroiled in consideration of the relative importance of pure recall, aided recall, recognition, visibility, prestige value, affective value, attention value, fixation time, confusion effect and so on, in the present study we took off from the simple question *"What does the advertiser want his advertisement to do?"*

For some the answer is that he wants to create an association in the consumer's mind between the *product* and his *brand name* To the man who has a new brand to sell an advertisement which leaves a brand impression on *more* people than competing advertisements is to that extent more successful

Therefore, in one test the respondent was exposed to advertisements in a setting which resembled the usual exposure of advertisements to magazine readers as closely as is possible in an interview situation This was followed by an interval to represent the period between the seeing of an advertisement and the occasion to

* Reprinted from *Journal of Applied Psychology* Vol 31 No 3, June 1947

buy the product During it the respondent's mind was taken off the advertisements he had seen, by questioning He could hardly have been deliberately attempting to retain the brand names in his mind because he was unaware that he would later be quizzed He could not have gathered the purpose of the study from the interviewers since they were uninformed Finally the respondent was read a list of products and asked to reply, if possible, with the brand names promoted by an advertisement he had been shown This process presumably represented the buying situation, where the customer feeling the need of a product does (or does not) call to mind a trade name

Experimental situations never perfectly mimic corresponding real life situations and this test is no exception It would have been better if the seeing of the advertise ments had been spontaneous rather than requested, if the interval had been longer and if the naming of a brand name had been, instead, the actual purchase (or failure to purchase) of a branded article Rec ognizing these and other limitations of the study, we must still admit that the test pits 4 color and black and white advertisements against each other on fairly equal terms We can look upon the data in hand as the basis for an estimate of the relative impact of these 4 color and black and white ad vertisements

But this test is particularly related to the function of only one kind of advertis ing A large part of the advertising in today s magazine is of trade names familiar to all Most people asked to name a gum will mention Wrigley s The word Ford is very definitely associated with an automobile The product brand name association could hardly be more firmly established What the advertiser must do with non-users of his brand is to create or increase the following association brand name prestige In other words, the advertiser must transform mere familiarity with the brand into real 'I want it With the people who already own and use his brand, the advertiser wants to create or preserve, and, if possible, to enhance this association *brand name—pride of ownership* The owner must be made to talk about his pos session to his less fortunate friends Also he must be made a repeat buyer

While the advertisement that aims to promote a new brand must primarily pound home the trade name, the adver tisement which seeks to heighten the luster of prestige and quality already associated with a familiar brand must, primarily, please attract, interest So the second of the two factors measured in the survey was the ability of an advertisement to intrigue or interest a reader This was done by asking the observers to indicate as they paged through a binder of full page advertisements any which particularly interested them

The Sample

For every 4 color advertisement promoting a given product used with one sub sample of 496 people, a corresponding black and white advertisement was used with a matching sub sample of 496 Thus, comparative measures of the two advertisements were obtained One thousand interviews were assigned but, actually, each sub sample was 4 interviews short

The one thousand were assigned in the manner indicated in Table 50 1 The obtained sample is given in Table 50 2 Education was unassigned but came out as shown in Table 50 3

The distribution of education was about what we would expect since we omitted the *D* economic group Obtained quotas for sex age and economic status approxi mated those assigned

The 10 cities were selected to provide a spread in type (industrial, market center, institutional) and in size

Los Angeles, California
Buffalo, New York
Urbana, Illinois
San Antonio, Texas
Altoona, Pennsylvania
Dover, New Jersey
Raleigh, North Carolina
Fargo, North Dakota
Chicago, Illinois
Hartford, Conn

A filter question eliminated from the

TABLE 50 1

Intended Dıstrıbutıon of Sample ın Each of Ten Cıtıes

Sample A Men	Sample A Women		Sample B Men	Sample B Women
		Sex and Age		
8	8	20–29	8	8
9	9	30–44	9	9
8	8	45 and over	8	8
25	25		25	25
		Economıc		
4	4	A	4	4
9	9	B	9	9
12	12	C	12	12
0	0	D	0	0
25	25		25	25

TABLE 50 2

Actual Dıstrıbutıon of Sample ın Each of Ten Cıtıes

Sub Sample A		*Sub Sample B*
	Age	
29 7%	20–29	31 0%
39 0	30–44	37 5
31 3	45 and over	31 5
	Sex	
47 7	Male	48 4
52 3	Female	51 6
	Economıc	
20 2	A	16 7
36 3	B	36 9
43 5	C	46 2
—	D	—
—	N A	2

TABLE 50 3

Obtaıned Dıstrıbutıon by Educatıon Level

Sub Sample A		*Sub Sample B*
6	N A No Schoolıng	2
11 0	Grammar School only	16 7
56 5	Hıgh School	52 3
31 9	College	30 8

sample all people who said they had not looked through a copy of one of the three leading general weeklies during the past 3 or 4 months The 992 people who replied Yes to this question were invited to go through a book in which were bound 20 full page advertisements from recent issues of a general weekly with very high circulation

SELECTION OF THE MATERIAL

Color and black and white should be represented by equally good advertisements Since the creators of advertisements are concerned with other matters than producing equivalent 4 color and black and white material for test purposes one might argue that it would be best to create synthetic advertisements for the purpose This, however, would be most unrealistic and would yield results of more interest to an academician than to a business man

In the present study, therefore, such uniformity in advertising excellency as was achieved derives from the assumption that there is a uniformly high level of creative ability applied to the production of full page advertisements recently appearing in a popular mass magazine having a very wide circulation

At first thought one might suppose that the highest degree of comparability would exist between two advertisements exactly alike in layout wording, picture, size, etc, and differing only in the dimension of color Actually we have good reason to believe that this is not true Experts in the field tell us that an advertisement created with only black ink in mind is inevitably different in many respects from one created for 4 color A layout of maximum effectiveness for the one would almost always fall short of maximum for the other It seemed wise for us to follow the experts in this matter and to select as most nearly comparable two advertisements prepared by the genius of a *single* advertising agency, advertising the *same* product in the *same* medium to promote the *same* trade name Therefore, these advertisements were selected arbitrarily as follows

Starting with the most recent issue of a mass weekly magazine and working backwards through four months, and in each case starting with the last page of an issue and working toward the front, *all* full page advertisements in 4 color or black and white were arranged in order The pairing of advertisements was determined by taking the most recent black and white and the most recent 4 color advertisement which promoted the same *product* (or service) and the same *trade name* Thus the experimenter exercised no judgment in the selection of an individual advertisement

The above procedure yielded 10 pairs of advertisements, which fulfilled our requirement Each pair consisted of a full page black and white and a full page 4 colored advertisement promoting the same product under the same trade name Here are the products

An electric lamp
A sheet
A cigarette
A whiskey
A railroad
A soap
A soft drink
An electric blanket
A wristwatch
A dentifrice

One member of each pair was bound in the folder used with half the sample, and the other in the folder used with the other half Five 4 color and 5 black and whites were included with each sample Further selections were made on the same arbitrary basis to provide two full page advertisements in 4 color and two in black and white advertising the same *product* (or service) but under *different trade names*

Five such quartets were found 4 makes of aeroplane, 4 makes of fabric for clothing, 4 makes of medicines, 4 makes of phonograph records, and 4 makes of cosmetics

Each half of the sample was tested on one black and white and one 4 color advertisement of each of the products Thus a total of 20 advertisements, half of them colored, was used with each respondent

Note that this selection was strictly arbitrary and automatically ruled out any prejudice or any flaw in judgment which might have influenced a selection made by the free choice of the investigator

The sequence of the 20 advertisements in the binder was arranged in accordance with the following rules

1 Black and white and 4 color alternated except that the regularity was broken by introducing two advertisements consecutively 4 color or black and white at five points in the series This was done so as to avoid the impression that black and white advertisements were being compared with 4 color

2 Advertisements promoting relatively costly, permanent acquisitions were irregularly mixed in with those promoting inexpensive quickly consumed items

3 Similar irregularity governed the sequence of items which were one of a pair and those which were one of a quartet Of the latter, the two belonging to the same quartet and, therefore, advertising the same product were separated by at least four intervening advertisements

Adjustment for previous exposure to advertising and thus possible variation in brand name familiarity was made in the ten cases where competing advertisements promoted different brands [1] This was done by the insertion in the following correction formula of the false mentions of a brand by respondents in that sample to which that brand was *not* shown

$$\text{Adjusted \% of sample correctly naming brand} = \frac{(\text{Obtained \% of correct mentions}) - (\text{\% false mentions})}{100\% - (\text{\% of false mentions})}$$

RESULTS

The study gives two measures of 4 color versus black and white The comparative tendencies of the two to arouse interest in an advertisement and their comparative tendencies to be recalled

Some respondents named only two or three advertisements as being interesting, some mentioned many Obviously the value of a mention depends upon the number so named We have, therefore, treated separately the number of respondents naming 0, 1, 2, 3, etc , up to all 20 interesting Because the distribution of the two samples among the groupings according to number of advertisements named as interesting is not identical, we weighted the raw figures in terms of the size of groups Thus we compared the group in Sample A naming 5 advertisements as interesting, with the group in Sample B naming 5 advertisements But we did so not in terms of *how many* in the one mention the black and white cigarette advertisement and *how many* in the other named the 4 color cigarette advertisement Rather, we used percentages in each case This was a defensible procedure, but it failed to yield a single index of the relative interest value of the two members of a pair To arrive at such a figure we first determined the similarity in judgment among the several groups and discovered a reasonably close relationship among respondents who name as interesting over three and fewer than 16 of the 20 advertisements The remaining groups behaved eccentrically We, therefore, combined the findings for the twelve groups naming from four to fifteen advertisements inclusive, by securing the algebraic sum of the differences in per cents naming the 4 color and the black and white advertisement in each pair

Actually, inclusion of the few individuals naming more than 15, or fewer than four advertisements does not appreciably alter the findings

The combined results are given in Table 50 4 In the first column is given the difference in per cent of people naming the 4 color and the black and white advertisement of each pair The base in every case is the number of people tested on the advertisement [2] In the second column is given

[1] This adjustment is the same as the confusion control used by D B Lucas

[2] In pairs marked A the 4 color member was shown subsample A and in pairs marked B the 4 color member was shown subsample B

TABLE 50 4

The Mean Dıfference ın Expressed Interest Between Black and Whıte and 4 Color and an Estımate of the Error of thıs Dıfference

	Mean Dıfference	Mean Dıfference ın Multıples of ıts σ*
A Dıfferences ın favor of 4 color		
Fabrıc A	36 6	11 35
Medıcıne A	33 4	11 72
Raılroad	31 8	7 91
Cosmetıc A	19 8	6 89
Blanket	19 7	7 87
Aeroplane B	18 0	3 79
Whıskey	15 5	4 91
Fabrıc B	14 6	2 94
Record A	13 1	4 06
Cosmetıc B	12 6	4 56
Medıcıne B	10 3	2 33
Record B	9 0	3 38
Cıgarette	8 8	3 71
Sheet	7 5	3 22
Dentrıfice	4 6	1 59
Soft drınk	3 3	1 23
Watch	3 3	16
B Dıfferences ın favor of black and whıte		
Electrıc lamp	3 6	1 59
Aeroplane A	4 1	1 73
Soap	8 2	3 04

* For each paır the dıfference ın the 12 groups (mentıonıng four, five, sıx etc) were averaged and the σ of these 12 dıfferences was computed Thıs σ was dıvıded by $\sqrt{n}$ to gıve an estımate of the σ of the mean dıfference Each mean dıfference was then dıvıded by the σ of that mean dıfference

the ratıo of average dıfference to the σ of the average dıfference, as an ındıca*ıon of the degree to whıch the weıghted dıffer ence, ıs tenable

The ımpact data were treated ın the same way Table 50 5 gıves the results

Obvıously when the black and whıte and colored members of a paır are com pared the latter, ın most cases, has the advantage ın both ınterest and ımpact value When judged by the ratıo of dıffer ence to error estımate, the advantage seems to be greater ın the case of ınterest One mıght ask, how does the factor stand out among the many others whıch are un doubtedly related to an advertısement's effectıveness?

A certaın degree of unıformıty exısts among the advertısements compared ın thıs study Any two promote the same product and were prepared by the same advertısıng agency In half the cases they promote the same trade name All were full page ad vertısements financed by firms whıch had bought space ın a magazıne wıth a very large cırculatıon and audıence, and whıch were, therefore under the same pressure to utılıze the space to advantage All ap- peared wıthın a 4 month perıod The only factor delıberately and constantly con- trasted was that of 4 color versus black and whıte Nevertheless, other factors, un- controlled and quantıtatıvely unıdentıfied, dıd vary In some cases one member of a paır had more text than the other In all cases the wordıng was somewhat dıfferent as was the ıllustratıon In a few the ap- peal was dıfferent It ıs possıble, ın the

TABLE 50 5

The Mean Difference in Expressed Impact Between Black and White and 4 Color and an Estimate of the Error of this Difference

	Mean Difference	*Mean Difference in Multiples of its σ**
A Differences in favor of 4 color		
Railroad	20 7	6 15
Fabric A	19 0	4 25
Medicine A	15 8	4 42
Cosmetic A	15 4	4 94
Record A	14 3	3 74
Cosmetic B	14 1	3 87
Fabric B	9 8	2 74
Cigarette	8 1	1 58
Electric lamp	7 8	2 66
Blanket	7 0	2 22
Record B	5 5	1 41
Aeroplane B	3 3	1 01
Dentifrice	1 8	61
Soap	1 0	42
B Differences in favor of black and white		
Sheet	4	14
Soft drink	1 0	39
Whiskey	4 6	1 47
Watch	5 0	2 21
Medicine B	10 0	3 29
Aeroplane A	11 3	3 95

* For each pair the differences in the 12 groups (mentioning four, five, six, etc) were averaged and the σ of these 12 differences was computed This σ was divided by $\sqrt{n}$ to give an estimate of the σ of the mean difference Each mean difference was then divided by the σ of that mean difference

case of half of the advertisements used (the quartets) to estimate the total effect of these uncontrolled factors upon the interest and recall values It will be remembered that each quartet consisted of two 4 color and two black and white advertisements, all promoting the same product We can hold the color factor constant by comparing the two 4 color members of each quartet with each other, and by comparing the two black and white members with each other

Table 50 6 presents the interest values and for comparison the 4 color versus black and white values are repeated

Judging by the ratios of mean difference to the σ of mean difference, we find that 4 of the 5 quartets contain a pair wherein the 4 color superiority is greater than the advantage of one or the other member in either case where color is compared to color, and black and white to black and white One quartet, aeroplanes, contains a black and white advertisement more superior to the other black and white than either difference in the color tests In interest value, then, color usually outweighted other factors

This is not true in the case of impact, however Impact values similar to interest values in Table 50 6 are given in Table 50 7

In impact values it seems clear that the standard differences are usually larger when color is compared with color and black and white with black and white than

TABLE 50 6

Comparison of Interest Indices Where Quartets were Used

4 Color versus Black and White			*4 Color versus 4 Color and b & w versus b & w*		
	Mean Difference	Mean Diff in Multiples of its σ		Mean Difference	Mean Diff in Multiples of its σ
Fabric A	36 6	11 35	Fabric, color	35 7	8 10
Fabric B	14 6	2 94	Fabric b & w	13 6	3 05
Medicine A	33 4	11 72	Medicine, color	28 6	10 13
Medicine B	10 3	2 33	Medicine, b & w	5 5	99
Cosmetic A	19 8	6 89	Cosmetic color	8 0	2 08
Cosmetic B	12 6	4 56	Cosmetic b & w	8	25
Record A	13 1	4 06	Record, color	5 5	1 68
Record B	9 0	3 38	Record, b & w	9 6	3 79
Aeroplane A	4 1*	1 73*	Aeroplane color	4 6	1 28
Aeroplane B	18 0	3 79	Aeroplane b & w	17 5	4 89

* Difference in favor of black and white All other differences in these two columns favor color

TABLE 50 7

Comparison of Impact Indices Where Quartets were Used

4 Color versus Black and White			*4 Color versus 4 Color and b & w versus b & w*		
	Mean Difference	Mean Diff in Multiples of its σ		Mean Difference	Mean Diff in Multiples of its σ
Fabric A	19 0	4 25	Fabric, color	22 4	6 41
Fabric B	9 8	2 74	Fabric, b & w	11 6	3 04
Medicine A	15 8	4 42	Medicine color	45 2	10 06
Medicine B	10 0*	3 29*	Medicine, b & w	20 3	8 65
Cosmetic A	15 4	4 94	Cosmetic, color	10 3	1 68
Cosmetic B	14 1	3 87	Cosmetic b & w	10 8	3 67
Record A	14 3	3 74	Record, color	17 4	5 64
Record B	5 5	1 41	Record, b & w	23 0	5 18
Aeroplane A	3 3	1 01	Aeroplane, color	3 9	11
Aeroplane B	11 3*	3 95*	Aeroplane b & w	18 4	7 81

* Difference in favor of black and white All other differences in these two columns favor color

they are when color is compared with black and white Apparently the other factors in combination exercise more influence than does color in the quartets It may very well be that this would be found true of the pairs also, had we a way to test the possibility

CONCLUSIONS

Obviously, the value of color in advertising depends upon a number of matters, such as the skill with which it is used, the adaptability of the product to black and white portrayal and so on

These tests indicate that a further consideration exists —the purpose of the advertiser They suggest that in the promotion of a new brand, the creation of association between product and trade name, color is not necessarily greatly superior In the protection of an investment in a familiar brand by keeping alive and increasing its reputation for quality, color appears to have a greater advantage over black and white It is possible that careful review of purpose in relation to the added cost of color may help to curb a trend toward uncritical selection of expensive presentation

The Use of the Free Association Technique in the Investigation of the Stimulus Value of Trade Names *

JOHN P FOLEY, JR

The free association technique, originally developed as a diagnostic method of revealing the idiosyncrasy *vs* the commonality of an individual's verbal associations, has been adapted for use in several fields of applied psychology, such as the measurement of associative reaction time, the classification of associations into various types, and the analysis of the subject s associations in order to reveal or diagnose interests, attitudes, guilty or technical knowledge, aptitudes, intelligence, emotional complexes, personality traits, pathological processes, and the like, as well as in psychotherapy The present study represents an attempt to use the free association technique in the investigation of the stimulus value of trade names,—a problem of considerable importance for the psychology of advertising, marketing, and related disciplines

SUBJECTS

The subjects employed in the present study were 300 George Washington University students, ranging in age from 16 to 46, with the majority falling between 18 and 21 The group was composed of approximately 40 per cent men and 60 per cent women The range in educational level was from Freshman to Post Graduate, although there was a predominance of Freshmen and Sophomores The subjects were divided into three experimental groups of 100 subjects each, the groups being equated with respect to age, sex, and educational level

PROCEDURE

The procedure employed was the standard free association method, although only one stimulus word was administered to each subject and only one response word was obtained When each subject had been provided with a pencil and blank 3′ × 5″ card, the following directions were given

"This is a brief test of verbal association I shall pronounce a word or name As soon as I have pronounced it, you are to write down on your card the *first* word that occurs to you, i e, the *first* word you think of after hearing the word or name that I

* Reprinted from *Journal of Applied Psychology,* Vol 28, No 5, October 1944

pronounce For example, if I say win dow and if you think of door, you would write down the word door imme diately

The word or name I give will be pro nounced only *once* it will not be repeated Do not converse with your neighbor, and do not look at his card We want your own individual reaction

Please remember you are to write down the very first word you think of after hearing the stimulus word Write only *one word* Write it down regardless of what it is

In one experimental group the stimulus word Pepsi Cola' was then given orally, after which each subject wrote down his verbal response The same procedure was followed in the other two experimental groups except that the stimulus words Coca Cola and Dixie Cola were given, respectively At the conclusion of the ex periment, each subject was asked to indi cate sex, age, and year in college

Results

In Table 51 1 will be found the complete alphabetized list of associations given to each of the three stimulus names In each case the subjects associations have been transcribed without change, even though the response was a phrase rather than a single word In the case of the responses Coca Cola and Pepsi Cola, abbrevia tions and colloquialisms as well as mis spellings have been classified separately under their respective names, although the total number of generic associations is also given in each case in parentheses Since there are 100 subjects in each of the 3 ex perimental groups, each entry in the first 3 response columns is a percentage, and intercomparisons can be made directly The last column represents the total num ber of associations made in the three ex perimental groups combined

It is interesting to note at the outset that there is a surprisingly small range of asso ciations to each of the three stimulus trade names Only 63 different responses were made by the 300 subjects, a result which would seem to indicate that associations to such trade names are limited in num ber and hence readily investigated by the free association technique The trade name Coca Cola seems to be associated with a slightly greater number of different things than the other two trade names, since there are 29 different responses to Coca Cola, 27 to Dixie Cola, and 22 to Pepsi Cola "

Let us next consider the extent to which a given stimulus trade name is associated with the trade name of another Cola prod uct as a response This is the most im portant part of the study as far as trade name confusion and infringement are con cerned If we consider the generic totals in making such a comparison, it is found that surprisingly large percentages of the total number of associations are of this type To the stimulus name Pepsi Cola," 22 per cent of the subjects gave the re sponse Coca Cola (or one of its generic equivalents), there are no other Cola" responses to this stimulus To the stimulus name Coca Cola," 10 per cent of the subjects gave the response Pepsi Cola," and 1 per cent gave the response R C Cola ' To the stimulus name Dixie Cola, 29 per cent of the subjects gave Coca Cola and 29 per cent gave Pepsi Cola, ' 1 per cent gave Clear Cola the only other Cola' response made in the present study [1]

From these figures, several interesting facts are clearly apparent Note that there are many more Coca Cola' responses to other Colas than there are other Cola" responses to 'Coca Cola, ' a result which would indicate that Coca Cola is the best known Cola product and that the subjects tend to think of it when they hear the name of another Cola product to a much greater extent than they tend to think of another Cola when they hear the trade name 'Coca Cola ' It will also be noted that the response ' Coca Cola is more frequent to the stimulus ' Dixie Cola than to the stimulus ' Pepsi Cola Similarly, the response Pepsi Cola is much more frequent to the stimulus Dixie Cola than

[1] The fact that no other Cola' responses occurred is of interest in the light of the extremely large number of differently named Cola products

TABLE 51 1

Response Word	*Stimulus Word*			
	Pepsi Cola	Coca Cola	Dixie Cola	Total
Awful		1		1
Bells	1			1
Beverage			1	1
Big	1			1
Bottle	11	22		33
Brown		1		1
Caffine		1		1
Candy	1			1
Canteen	1			1
Clear Cola			1	1
Coca Cola (generic)				
Coc		1		1
Coca			1	1
Coca Cola	16		26	42
Coke	6		2	8
(Total generic)	(22)	(1)	(29)	(52)
Cola			1	1
Cold		2	1	3
Commercial			1	1
Dixie cup			1	1
Drink	40	32	11	83
Drinking		1		1
Drug store		2		2
Excitement	1			1
Flying	1			1
Fountain		2		2
Ginger ale		1		1
Glass		3		3
Good	1			1
Hits	4			4
Hits the spot	3			3
Ice		1		1
Ice cream			5	5
Large bottle	1			1
Machine		1		1
Magnolia tree			1	1
Mint		1		1
Music—swing			1	1
Nickel (nickle)	3	1		4
Paper cup			1	1
Pensa cola			1	1
Pensicola, Florida			1	1
Pepsi Cola (generic)				
Pepi Cola			1	1
Pepsi Cola		10	27	37
Pespi Cola			1	1
(Total generic)		(10)	(29)	(39)
Peter Arno	1			1
Pop	1	2	2	5
R C Cola		1		1
Red			1	1
Reddish brown		1		1
Refreshes		1		1
Refreshing		1		1

TABLE 51 1 (*Continued*)

Response Word	*Stimulus Word*			
	Pepsi Cola	Coca Cola	Dixie Cola	Total
Refreshment(s)		3	1	4
Root beer		1		1
Sandwich		1		1
Sherbet			1	1
Soda	2	3		5
Soft drink	2		4	6
South			4	4
Sparkling beverage			1	1
Straw		1		1
Tastes	1			1
Thirst	1		1	2
Vanilla		1		1
Water	1	1		2

to the stimulus Coca Cola' In fact the responses of Pepsi Cola and of Coca Cola to the stimulus Dixie Cola are of equal strengths (29 per cent) Dixie Cola, it is found, is not given as a response to either Pepsi Cola or 'Coca Cola On the basis of these results it would appear that Pepsi Cola is a stronger competitor of Coca Cola than is Dixie Cola, since it is sometimes given as a response to the stimulus 'Coca Cola" (10 per cent) and at the same time 'Coca Cola is not as frequently given as a response to it (22 per cent) as to Dixie Cola (29 per cent) [2] The fact that Dixie Cola' is not given as a response at all would tend to strengthen this conclusion

Let us now consider the remainder of the associations, i e, those which do not involve specific Cola trade names as responses The most common of these responses is 'Drink,' this response being given in 40 per cent, 32 per cent, and 11 per cent of the cases in the Pepsi Cola," Coca Cola, and 'Dixie Cola groups, respectively The relatively smaller incidence of this association in the last named group is tied up with the fact that 58 per cent of the responses of this (Dixie Cola) group were to the trade names Coca Cola" and Pepsi Cola The only other non trade name response which appears with great frequency is Bottle,' being given in 11 per cent of the cases in the Pepsi Cola stimulus group and 22 per cent of the cases in the Coca Cola stimulus group, it is interesting to note that this response does not occur at all in the Dixie Cola stimulus group The greater incidence of this response to the stimulus word 'Coca Cola undoubtedly reflects the greater familiarity and perceptual distinctiveness of the Coca Cola bottle

The remaining non trade name associations occur with small and scattered frequencies, and often fall into certain general categories Noteworthy are those involving *color* ('Brown, Red, Reddish brown), *generic terms* (Beverage,' Pop, Soda, Soft drink,' Sparkling beverage'), *size* (Big," 'Large bottle'), *associated contiguous objects* (Candy," 'Canteen, Dixie Cup, Drug store," 'Fountain, Ginger ale Glass Ice,' Ice cream,' 'Machine,' Mint, Paper cup, Root beer," Sandwich, Sherbet, Straw) *effects* (Refreshes,' Refreshing'), *evaluation* ('Awful,' Good), and sound association (Pensa Cola, Pensi cola Florida —both given to Pepsi Cola') *Geographical associations* are involved in the 4 responses of South and single response of Magnolia to the stimulus Dixie Cola The *influence of radio and other forms of advertising* is also indicated by the responses 'Big, Canteen,"

[2] It would be interesting to study the effect of advertising in setting up a differentiation between two such products on the part of the consumer public

'Hits," "Hits the spot," "Large bottle," 'Nickel," and Peter Arno' to the trade name Pepsi Cola," and by the responses Refreshes' and Refreshing to the stimulus name Coca Cola In this latter connection it is of interest to note that the response Thirst is given by only 2 subjects and does not occur to the stimulus Coca Cola, in spite of the advertising slogan, Thirst knows no season'

The Psychogalvanometric Method for Measuring the Effectiveness of Advertising *

* The authors are indebted to Mr G Maxwell Ule, Director of Research, McCann Erickson, Inc, Chicago, Ill, for permission to use the ads and appeals used in this study and for the sales test results used as a criterion in this study

GORDON ECKSTRAND and A R GILLILAND

Advertisers have long been searching for objective techniques or methods of pre testing advertising material which are inexpensive, fast and reasonably valid That is, a technique or method of predicting, in advance of use in an advertising campaign, the effectiveness of certain advertising material as judged by a criterion of volume of sales induced

Whether an advertisement is a good one or not can only be determined, in the last analysis, by running the ad as scheduled and then observing the effect on sales exclusive of other factors The buying public is, after all, the final judge But this is an expensive method of operating considering both time and money, since it does not permit the weeding out of poor ads before they are put before the public as part of an advertising campaign In 1946 more than two billion dollars was spent for all kinds of advertising With this great amount of money being spent, it is important for advertisers to get as much as possible out of each advertising dollar Thus the pre testing of advertisements is of great economic interest as well as an interesting problem in the prediction of human behavior

(In an attempt to get some idea of what to expect from an advertising appeal in advance of its actual use on the public, and in an effort to determine what factors go toward making good and poor ads, advertisers have developed several techniques for testing their material Experts judgments, cross sections of public opinion, point rating systems, memory for ads, point of purchase sales tests, and split runs in media of limited circulation have all been used to test advertising material However, some of these techniques have shown little validity, and others are time consuming and costly Consequently the field of advertising is still looking for a valid and rapid method of measuring the effectiveness of advertising matter)

It is the purpose of this research to investigate the usefulness of the psychogalvanic response as a measure for use in predicting the effectiveness of advertising material as measured by a sales test criterion

(For a good many years after its discovery as a psychological measuring tool in 1888, the psychogalvanic phenomena enjoyed almost unbelievable popularity in psychological research) It has been studied with reference to everything from attitude (1) to the effect of cobra venom (7) However, when we turn to a consideration of the psychological correlates of the psychogalvanic response we find little agreement among investigators At various times and by various investigators, the psychogalvanic response has been claimed as a

* Reprinted from *Journal of Applied Psychology* Vol 32, No 4, August 1948

measure of emotion, conation, attitude, attention, level of consciousness, and many others (5).

Landis and Hunt (6) have pointed out that the galvanic response is not a measure of, regular criterion of, or indicator of, any one or a combination of these traditional psychological categories. However, as both Landis and Darrow (3) have agreed, it seems to be a fairly certain method of demonstration of general autonomic activity.

It seems fairly well established, then, that while many stimuli and stimulus situations may serve to elicit the psychogalvanic response, the response seems to be a good measure of the amount of general bodily arousal present at any time or during any portion of behavior. It seems equally well established that the psychogalvanic response is not a valid and reliable measure of any of the traditional psychological categories. This does not necessarily mean, however, that the psychogalvanic response will not be of value in predicting certain more complex types of responses. It may be that in a response as complex as a person's reaction to an advertisement, several or many of the psychological conditions mentioned above may be present and affecting behavior. It is this total response to the situation, this total amount of arousal in which we are interested. The psychogalvanometer seems well suited to measure this total arousal.

There has been very little work done using the galvanometer to test advertising material. However, some evidence has accumulated to indicate that the changes in skin resistance of selected samples of subjects exposed to advertising material may be of value in predicting the later effectiveness of that material. Ruckmick (8) conducted a study in which the responses of the sweat glands were recorded during a three second exposure of advertising copy. Several series of copy, run with twenty subjects, revealed an internal consistency of data and also gave results which tallied in a general way with the choices obtained by the serial procedure of impression.

Conrad (2) [1] conducted an investigation to determine whether it was possible to study the responses made to advertising appeals of car cards by means of a psychogalvanic response apparatus. Using a Hathaway galvanometer, he exposed a series of car cards to a large group of subjects for five seconds each. The subjects used were college students and the cards were presented in a counterbalanced order. He later had the subjects rank the ads as to their effectiveness in getting attention. He found that the results obtained in this manner correlated only .18 with the results obtained by the galvanometer. He did find, however, that definite galvanic responses could be obtained with advertising material as stimuli, and that certain material got larger responses than other material.

Gilliland and Sharp (4) showed that the psychogalvanometer does record variations in the effect of advertising on readers. They did not attempt, however, to correlate the size of the subjects' galvanic reactions with the effectiveness of the ads as determined by an outside criterion. They pointed out the need for using the psychogalvanometer to test ads that had already been evaluated as to selling effectiveness in order to establish the validity of the method.

In these earlier studies the technique has not been subjected to a rigid experimental test where a suitable subject group was used and where the method was validated against a suitable objective criterion. The few studies reported here have used either no criterion of the ads' effectiveness or have used only the subjects' opinion. This is, at best, only a criterion of very limited value. The best, most direct, and most objective criterion readily available is some measure of the ads' actual selling effectiveness in a realistic advertising situation. It is the purpose of this research to test the hypothesis that effective advertising material, as judged by a sales criterion, will, on the average, induce larger psychogalvanic responses in a selected sample of the population than will less effective advertising material.

---

[1] This investigation was done under the direction of Dr. A. R. Gilliland.

THE EXPERIMENT

Subjects The material tested dealt with three popular nationally advertised food products made by the same company An attempt was made to obtain a subject sample which would approximate a sample from the population to which the ads were directed Since the material dealt with nationally advertised products, the sample used falls short on one count immediately The sample used had to be drawn from the area in and around Evanston, Ill Evanston and the surrounding area cannot be considered a representative section of the country, but the sample drawn from this area seems more representative of the country at large than it does of the Evanston area

Since the material dealt with in this study was concerned with basic food products, the sample was made up of married women or single women who cook and purchase groceries A few women were included who were engaged to be married and thus will soon be part of the potential buyers of these products An attempt was made to get a distribution of subjects from the various income and age groups and a distribution of subjects with and without children Due to the difficulty of obtaining subjects, no attempt was made to match local or national statistics on these factors Table 52 1 presents the number of subjects falling in each of the categories

TABLE 52 1

Analysis of the Subject Group

Income	*Number*	*Age*	*Number*	*Children*	*Number*
Below $3,000	18	Below 24	15	No children	29
$3,000–$5,000	16	25–39	18	Children	19
Above $5 000	14	40–54	11		
		Above 55	4		

Ads and appeals used Three series of advertising material were tested Two of the series consisted of advertising appeals made up into finished advertisements and the third series was composed of advertising appeals in verbal form not yet made up into ads

Series 1 consisted of three finished ads of pancake flour Each ad was 11′ by 8½ and was done in black and white With respect to all variables but basic appeal the ads were quite similar They contained about equal amounts of pictorial illustration, headlines of approximately equal length, about the same amount of copy, and the brand name was used equally often Series 2 consisted of two finished ads dealing with a baby food Each ad was 16 by 9 and was done in black and white Again the ads were quite similar with respect to all variables but basic appeal All the finished ads were mounted on stiff, white cardboard Series 3 was made up of four advertising appeals of themes of a popular brand of flour These were basic themes or ideas which might later be used as a basis for the formulation of finished ads Since the sales test to be used as a criterion was made with verbal presentation of the appeals it was decided to record the appeals so that they could be presented to the subjects in a similar manner The appeals were recorded by an announcer with radio experience The announcer was told to make each presentation as constant as possible He was informed as to the nature of the experiment and told that we were interested in measuring the effectiveness of the basic theme or idea contained in the appeal and did not want effectiveness to vary as a function of the different qualities of his presentation It is not possible to ascertain how well this purpose was accomplished, but of the several persons who have listened to the presentations, none have detected any bias in favor of any one appeal

Procedure Two rooms were used in this investigation One room was used for the presentation of the advertising material to

the subject, and the adjoining room contained the equipment for recording the psychogalvanic responses and the equipment for playing the recorded material One experimenter was in the room with the subjects and gave instructions and presented the material The other experimenter was in the adjoining room and controlled the recording apparatus The experimenters were in contact with each other by means of a two way signal system

The room in which the subject was seated was bare of distracting influences as far as this was possible The room was semi sound proofed, and although it did not keep out all sounds, it reduced the extraneous noises to a minimum All day light was excluded and the room was lighted by electric lights so that the light on the ads would be constant The ads were presented on a stand which was adjustable for height and distance and were presented at eye level A blank piece of white cardboard covered the first ad and a similar piece separated each of the following ads so that the experimenter could control the rate of presentation

When a subject arrived she was brought into the room, and the electrodes were fastened to her palm and arm As most people have a distinct aversion to being shocked by an electric current, this disturbing influence was removed as far as possible by telling the subject that there was absolutely no danger of being shocked The subject was told to sit relaxed and that all that was required of her was to look at and listen to the material as it was presented She was told to look at the ads as if she were seeing them in a newspaper or magazine and to listen to the appeals as if she were hearing them over the radio or someone was saying them to her

Within any series, the ads and appeals were presented in a counterbalanced order, and the presentation of the series themselves was also counterbalanced This procedure controlled position effects and the inter and intra series influences of an ad or appeal on another

The subject was allowed to relax for a period of three to five minutes after the completion of the instructions in order for her to get used to the situation This tended to make the galvanic readings more stable At a signal from the experimenter running the apparatus, the other experimenter removed the first blank card thus exposing the first ad In order to accustom the subject to this procedure the first printed advertisement and recorded appeal were always dummies during which time no readings were taken This also tended to make the galvanic readings more stable The ads were presented for a 30 second period while the appeals lasted about 15 seconds Between 30 and 45 seconds was allowed between the presentation of the ads and appeals within a series and between 45 and 60 seconds was allowed between each series This interval depended upon the stability of the psychogalvanic readings at the time The beginning and end of each exposure period was marked on the tape recording of the subject s responses

Apparatus The apparatus used in obtaining the galvanic readings was a two stage vacuum tube voltage amplifier with direct coupling It was designed specifically for this type of research and this type of measurement The apparatus has the advantage of ease of manipulation, accuracy in giving quantitative comparisons, and high sensitivity An additional advantage was the obtaining of permanent records by graphically recording the psychogalvanic responses by means of an Esterline Angus graphic recorder model A W

Zinc electrodes about one inch in diameter were used These were attached by means of leather straps and sponge rubber between the electrode and the strap assured an even contact with the skin area One electrode was attached to the palmar surface of the hand and the other to the inner surface of the forearm Commercial electrode paste and jelly were used to facilitate contact with the subject s skin area

The graphic chart of the recorder moved with a speed of three inches per minute and the magnified changes in the subject circuit were recorded on the moving chart by means of a writing mechanism The machine was calibrated with a decade resistance box so that the recorded

responses could be read off as changes in subject resistance

Criterion [2] The criterion used in all three series of ads and appeals was the results from sales tests conducted by the McCann Erickson Advertising Agency in Chicago The purpose of these sales tests was in each case to analyze the relative sales effectiveness of the ads and appeals in question

The tests, in each case, were made through a study of the movement of store inventories associated with consumer exposure to the alternative advertising material studied The studies were all conducted using stores situated in what were believed to be representative urban communities In the consumer exposure to the various advertising materials, strict counterbalancing techniques were used This tended to control the effect of random factors, biases from the cumulative impact of advertising exposure, and from the sequence of presentation of the various appeals

In all of the tests strict and rigid controls were used, therefore, since advertising was the major variable in the stores during the test, it is reasonable to assume that the differences in sales, revealed by the store inventories, was the result of advertising

Of the many possible ways of evaluating the changes in resistance, only one was used in this study—the total log conductance change during the exposure to any ad That is, the log conductance change for each ad was summated giving a total arousal value However no change was recorded unless there was at least 200 ohms of change and no differences between ads were recorded unless the change was 10 per cent or greater

Results

The problem of this study was the relationship between the total arousal produced by the ad and its sales effectiveness If two ads had equal arousal they would produce equal log conductance changes or one would be greater in half of the cases and the other would be greater in the other half Any variation from this one to one relationship could reasonably be attributed to the greater efficiency of one appeal over the other [3] The significance of any deviation from this ratio can be checked by the Chi square method Table 52 2 gives the number of times each ad in each of the three series produced the largest arousal value and the Chi square values for these differences

From Table 52 2 we can examine each of the three series of ads In the pancake flour ads it is apparent that ad A gave more high arousals than ad B The chi square value of 3 26 would occur by chance not more than about seven times out of 100 The chi square of 3 78 between A and C would occur not more than about five times out of 100 The difference between B and C was insignificant In the baby food series there were no significant differences between the two ads There were likewise no significant differences in the flour ad appeals

These same data for the arousal value of the three series of ads were analyzed by another method The smallest log conductance obtained for each subject was arbitrarily given a value of 0 and the highest value obtained a value of 10, other values were distributed between these extremes Table 52 2 gives the means for each ad in each series by this method

The difference between these means were checked for significance by the Fisher 't test Table 52 4 gives the t value for each comparison for each of the three series of ads

The t score between ads A and B for the pancake flour ads was 1 60 This means that if no difference in effectiveness existed between the ads a "t as large as this and in the same direction would be obtained in only about seven times out of 100 by

[2] A more complete description of the criterion tests cannot be given due to the confidential nature of the techniques

[3] The authors are aware that other assumptions can be made about the distribution of the expected frequencies and the treatment of the cases in which no differences were found in galvanometric readings between the ads in a series However, any method of calculation would give similar results and the method here used seems as defensible as any

TABLE 52 2

Differences in Effectiveness of Three Series of Ads

Comparisons Made	Pancake Flour Observed Frequency	Pancake Flour Expected Frequency	Pancake Flour Chi square	Baby Foods Observed Frequency	Baby Foods Expected Frequency	Baby Foods Chi square	Flour Observed Frequency	Flour Expected Frequency	Flour Chi square
A = B	7			9			23		
A > B	24	18 5	3 26	18	19 5	24	11	12 5	36
B > A	13	18 5		21	19 5		14	12 5	
A = C	8						24		
A > C	25	19	3 78				12	12	00
C > A	13	19					12	12	
B = C	10						26		
B > C	18	18	00				11	11	00
C > B	18	18					11	11	
A = D							19		
A > D							13	14 5	32
D > A							16	14 5	
B = D							23		
B > D							13	12 5	04
D > B							12	12 5	
C = D							23		
C > D							13	12 5	04
D > C							12	12 5	

TABLE 52 3

Mean Reactions for Each Ad

Ad	Pancake Flour N	Pancake Flour Mean	Baby Food N	Baby Food Mean	Flour N	Flour Mean
A	46	5 50	48	3 73	48	1 31
B	46	4 03	48	4 52	48	1 49
C	46	3 98			48	1 28
D					48	1 68

chance The t value of 06 between B and C was insignificant

The difference between the baby food ads would occur about 17 times out of 100 by chance and was therefore on the border line of probable significance None of the flour ads showed statistically significant differences

Both of the above types of analysis lead to the same general results The results for the two methods can now be compared with the sales efficiency of the ads as a measure of the value of the galvanometric method of testing ads

Criterion data In the sales test conducted with the pancake flour ads it was found that ad A sold 2 1 times as many packages of flour as did either of the other two ads There was little difference between ad B and ad C Ad A sold 100 units of flour, ad B 47 units, and ad C 48 units

In the sales test on the baby food ads, no significant difference was found in the selling effectiveness of the two ads Ad A sold 92 units and ad B sold 100 units

In the sales test conducted using the four advertising appeals or themes, it was concluded that there was a significant difference in the relative sales effectiveness of the four appeals tested The differences were small, however, and the advertising agency concluded, that for practical advertising purposes, the actual degree of difference was so small that any of the appeals could be used with about equal effectiveness Appeal A sold 96 units to the people hearing its sales talk, appeal B sold 100 units, appeal C sold 83 units and appeal D sold 90 units

Summary

Close agreement was found between the galvanic changes produced by a series of pancake ads and the sales effectiveness of these ads The sales effectiveness of ad A was 2 1 times as great as either ad B or C Little difference was found between B and C Both the Chi square method and the "t" test indicated that ad A was better (galvanic responses) than the other two ads at the 7 per cent level of significance By the method described here, no attempt was made to determine how much A exceeded B and C B and C were not significantly different in their galvanic responses

The baby food ads had almost equal sales appeal In their galvanic responses there was no statistically significant difference

TABLE 52 4

t Test for Significance of Difference

Comparison	Pancake Flour *t*	Baby Food *t*	Flour *t*
A–B	1 60	90	33
A–C	1 53		06
B–C	06		48
A–D			58
B–D			37
C–D			73

The results are more equivocal for the four flour ads, although the sales tests showed statistically significant differences These differences, however, were small and the advertising agency stated that for practical purposes the four appeals could be considered equal The differences between the galvanic responses to these appeals were not statistically significant

(In conclusion, it may be stated that this study adds positive evidence in behalf of the hypothesis that, under properly controlled conditions, the effectiveness of advertising material can be predicted by the psychogalvanic method[4] Further work is needed, of course, with different types of advertising material and with material of different degrees of effectiveness However, the technique gives promise as an objective evaluation of ads and advertising appeals)

[4] This statement is supported not only by the experimental results reported here but also by a series of extensive but less carefully controlled studies briefly summarized in a popular article by Walter P Wesley Results of Copy Testing by Arousal Method ' *Advertising and Selling* Nov 1947

REFERENCES

1 Abel, T M Attitudes and the Galvanic Skin Reflex *Journal of Experimental Psychology* 1930 Vol 13, 47-60

2 Conrad W E F *The Effect of Advertising on Psychogalvanic Reactions* Unpublished thesis Northwestern University, 1929

3 Darrow, C W The Equation of the Galvanic Skin Reflex Curve I The Dynamics of Reaction in Relation to the Excitation Background *Journal of General Psychology* 1937 Vol 16 285-309

4 Gilliland, A R, and Sharp L H Unpublished study

5 Landis C Psychology and the Psychogalvanic Reflex *Psychological Review* 1930, Vol 37 381-398

6 Landis, C and Hunt, W A The Conscious Correlates of the Galvanic Skin Response, *Journal of Experimental Psychology* 1935, Vol 18 505-529

7 Macht D I, and Macht M B Effect of Cobra Venom and Alkaloids on the Psychogalvanic Reflex *American Journal of Physiology* 1940, Vol 129 412

8 Ruckmick C A The Electrodermal Response to Advertising Copy' *Psychological Bulletin* 1939, Vol 36 627

PART FIVE

Newer Concepts

Industrial psychology must be viewed as an ever expanding body of knowledge Section III indicates the growth of a part of the field since World War II and its assimilation into a more or less distinct subject matter Ordinarily growth in industrial psychology does not take place in such a manner but rather through the introduction of newer concepts as modifications or improvements upon older and accepted ideas Ghiselli [1] reporting on new ideas in Industrial Psychology chose to mention six, namely, Lewin's concepts related to motivation of workers, Likert and Katz's studies on workers' morale, Shartle's approach to leadership, Haires' phenomenological attack on the problem of industrial peace, Flanagan's critical requirements, and Thorndike's formulation of personnel classification These points illustrate the usual growth based upon modification and improvement of existing techniques or concepts

It was deemed advisable to devote one section of this volume to newer concepts The manner of selecting the specific concepts was purely a matter of judgment but the three chosen were considered as both provocative and promising The Flesch Formula, Forced Choice, and Critical Requirements are presented to enable the reader to become familiar with the steps involved in using the tech niques as well as to become familiar with some of their applications It is im portant to curtail the lag in time between the proposal of a concept and its evaluation and widespread use In psychology, techniques and concepts cannot be kept secret and if they are used in guarded fashion by only the sponsors then evaluation is extremely limited if possible at all

Chapter XIII

THE FLESCH FORMULA AND SOME APPLICATIONS

A virtual hornets' nest has been stirred by Dr Rudolph Flesch and his readability yardstick Essentially, it consists of two formulas Formula A is a measure of abstraction or reading ease It considers sentence length in words and average word length in syllables Formula B is a measure of human interest It considers the percentage of "personal words" and "personal sentences "

Application of these formulas raises such basic questions as who can read and

[1] E E Ghiselli New Ideas in Industrial Psychology *Journal of Applied Psychology* 1951, Vol 35, 229–235

who can write. It also allows for considerable emotional expression concerning stylized writing or art. Be that as it may, more and more reference will be made to this contribution by Flesch, and many unexpected applications of his yardstick will cause sputtering for some time to come in unexpected and various places.

The formulas are simple to understand and to apply, as the Flesch article in this section clearly indicates. The Hayes, Jenkins, and Walker paper is an example of subjecting the system to a test to determine its reliability. Obviously, a measuring instrument without at least the characteristic of reliability is not a scientific measuring instrument. The results indicate that different people obtain similar results regardless of previous training. Pashalian and Crissy have applied these formulas to corporate annual reports, finding the reports difficult and dull. Farr, Paterson, and Stone apply the formulas to a sample of management and union publications. Such publications fall short of their mark since their level of difficulty is too high and their human interest is too low.

A New Readability Yardstick *

RUDOLF FLESCH

Samples from the main body of this paper, when tested for readability by the method here proposed, had an average 'reading ease' score of 30 and a 'human interest' score of 0. Presumably, the paper is easier to read than most other articles appearing in scientific journals. The section, 'The Formulas Restated,' which contains directions for users of the formulas, has a 'reading ease' score of 79 and a 'human interest' score of 42—which puts that portion of the article in the class of a good cookbook.

In 1943 the writer developed a statistical formula for the objective measurement of readability (comprehension difficulty) (5, 6). The formula was based on a count of three language elements: average sentence length in words, number of affixes, and number of references to people. Since its publication, the formula has been put to use in a wide variety of fields. For example, it has been applied to newspaper reports (9, 20), advertising copy (1), government publications (19), bulletins and leaflets for farmers (3), materials for adult education (4), and children's books (12). Its validity has been reaffirmed by 5 independent studies: the formula ratings of psychology textbooks substantially agreed with ratings by students and teachers (17), the formula scores rated specially edited radio news, newsmagazine, and Sunday news summary copy 'more readable' than comparable newspaper reports (18), advertisements, rated 'more readable' by the formula, showed higher readership figures (7), and articles that were simplified with the aid of the formula brought increased readership in two successive split run tests (13, 14). Since 1943, a number of academic institutions have incorporated the formula in the curriculum of courses in composition, creative writing, journalism, and advertising; it has also been used as the basis of several graduate research projects.

Because of this wide application, it seemed worthwhile to re-examine the formula and to analyze its shortcomings. One of these is to be traced to the basic structure of the formula; others are the results of difficulties in its application.

The structural shortcoming of the formula is the fact that it does not always

* Reprinted from *Journal of Applied Psychology*, Vol. 32, No. 3, June 1948.

show the high readability of direct, con versational writing For example, in the study of psychology texts mentioned above (17), the score of Koffka s *Principles of Gestalt Psychology* (the students choice for unreadability') was 5 4 ('difficult), yet William James *Principles of Psychol ogy* a classic example of readability, rated 6 0 (bordering on 'very difficult) Simi larly, the formula consistently rates the popular *Reader s Digest* more readable than the sophisticated *New Yorker* mag azine although many educated readers consider the *Reader's Digest* dull and the sprightly *New Yorker* 10 times as readable

Aside from that, the practical applica tion of the formula led to several minor misinterpretations Sentence length for in stance, is the element with the heaviest weight, it is also the easiest to measure As a result, this feature of the formula is often overemphasized, sometimes to the exclu- sion of the others—as in the directives that have been issued to staff writers of the Associated Press and the *New York Times* recommending the use of shorter sentences in leads On the other hand, the second element—number of affixes—seems often difficult to apply, users of the formula found this count particularly tedious and admitted to uncertainty in spotting affixes The third element—references to people—raised no such questions, but it was some times felt to be arbitrary and the under lying principle was often misunderstood

In addition many people found it hard to get used to the scoring system, which generally ranges from 0 (very easy) to 7 (very difficult") Also, the average time needed to test a 100 word sample is six minutes (4) This makes the application of the formula considerably faster than that of earlier formulas, which required reference to word lists (e g, Gray Leary (8) or Lorge (10)), but it is still too long for practical use

The revision of the formula presented in this paper is an attempt to overcome these shortcomings and make the formula a more useful instrument

Procedure

The criterion used in the original for- mula was McCall Crabbs *Standard Test Lessons in Reading* (11) The formula was so constructed that it predicted the aver age grade level of a child who could answer correctly three quarters of the test ques tions asked about a given passage Its mul tiple correlation coefficient was $R = 74$ It was partly based on statistical findings established in an earlier study by Lorge (10)

For many obvious reasons, the grade level of children answering test questions is not the best criterion for general read ability Data about the ease and interests with which adults will read selected pas sages would be far better But such data were not available at the time the first formula was developed, and they are still unavailable today So McCall Crabbs *Standard test lessons* are still the best and most extensive criterion that can be found, therefore they were used again for the re vision

In reanalyzing the test passages, the fol lowing elements were used

(1) *Average sentence length in words* The same element was used in the previous formula, but the correlation coefficient used was taken from Lorge s earlier find ings In the present study this coefficient was recomputed

(2) *Average word length in syllables* expressed as the number of syllables per 100 words The hypothesis was that this measure would furnish results similar to the affix count in the earlier formula Syl lables are obviously easier to count than affixes since this work can be reduced to a mechanical routine

(3) *Average percentage of "personal words '* The same element was used in the earlier formula However, the opportunity was used to test a clarified definition, which made no significant difference in correlation The new definition was stated as follows All nouns with natural gender, all pronouns except neuter pronouns, and the words *people* (used with the plural verb) and *folks*

(4) *Average percentage of "personal sentences"* This new element was designed to correct the structural shortcoming of the earlier formula, mentioned above By hypothesis, it tests the conversational qual

ıty and the story ınterest of the passage analyzed It was defined as the percentage of the followıng sentences Spoken sen tences, marked by quotatıon marks or otherwıse, questıons, commands, requests, and other sentences dırectly addressed to the reader, exclamatıons, and grammatıcally ıncomplete sentences whose meanıng has to be ınferred from the context

To make the predıctıon more accurate, 13 of the 376 McCall Crabbs passages that contaıned poetry or problems ın arıth metıc were omıtted ın the count of the first two elements, whıch are desıgned to test solely prose comprehensıon However, these 13 passages were retaıned ın the count of the last two elements, whıch are desıgned to test human ınterest

Followıng the procedure ın the earlıer study, ıntercorrelatıons were then com puted However, multıple correlatıon of the four elements wıth the crıterıon showed no sıgnıficant gaın ın predıctıon value over the earlıer formula ın spıte of the sıgnıfi cant predıctıon value of the addıtıonal fourth element by ıtself ($r = - 27$) Therefore, *two* multıple correlatıon regres sıon formulas were computed one usıng the first two elements and one usıng the last two Thıs procedure had the advantage of gıvıng ındependent predıctıons of the readıng ease and the human ınterest of a gıven passage

Fınally, the resultıng twın formulas were expressed ın such a way that maxımum readabılıty (ın both formulas) had a value of 100, and mınımum readabılıty a value of 0 Thıs was done to make the scores more readıly understandable for the prac tıcal user

FINDINGS

The ıntercorrelatıons, means, standard devıatıons, and regressıon weıghts found are shown ın Tables 53 1, 53 2, and 53 3 The followıng symbols were used *wl* for word length (syllables per 100 words), *sl* for sentence length ın words, *pw* for per centage of "personal words," *ps* for per centage of 'personal sentences," C_{50} for the average grade of chıldren who could an swer one half of the test questıons cor rectly, and C_{75} for the average grade of chıldren who could answer three quarters of the test questıons correctly

The two regressıon formulas based on these correlatıons are

Formula A (for predıctıng "readıng ease') $RE = 206\,835 - 846\ wl - 1\,015\ sl$

The scores computed by thıs formula have a range from 0 to 100 for almost all samples taken from ordınary prose A score of 100 corresponds to the predıctıon that a chıld who has completed fourth grade wıll

TABLE 53 1

Correlatıons, Means, Standard Devıatıons, and Regressıon Weıghts of Word and Sentence Length

	sl	C_{50}	$\overline{X}$	*s*	*B*
wl	4644	6648	134 2208	13 6845	5422
sl	——	5157*	16 5213	5 5509	2639

* After the preparatıon of thıs paper two artıcles appeared that poınted out a computa tıonal error affectıng the wrıter's orıgınal formula (E Dale and Jeanne S Chall, A Formula for Predıctıng Readabılıty,' *Educatıonal Research Bulletın* Ohıo St Unıv, 1948, Vol 27 11–20, 28, Lorge, I The Lorge and Flesch Readabılıty Formulae a correctıon *Sch & Soc*, 1948, Vol 67, 141–142) The error concerned the correlatıon coefficıent between sentence length and the crıterıon, whıch had orıgınally been reported by Lorge as 6174 the wrıter, acknowledgıng hıs debt to Lorge, used that figure wıthout recomputatıon The corrected correlatıon coefficıent ıs now reported as 4681 by Dale and Chall and as 467 by Lorge, thıs corresponds closely to the figure of 5157 reported ın Table 53 1, con sıderıng the fact that the wrıter now used a slıghtly better crıterıon of 363 passages for sentence length In other words, the formula presented ın thıs paper ıncıdentally and ındependently also corrects the error found by Dale and Chall and by Lorge

TABLE 53 2

Correlations, Means, Standard Deviations, and Regression Weights of Personal Words and Sentences

	ps	C_{50}	$\overline{X}$	s	B
pw	2268	− 3881	7 3457	5 5175	− 3446
ps	——	— 2699	29 5745	35 5822	— 1917

TABLE 53 3

Means and Standard Deviations of Two Criteria

	$\overline{X}$	s
C_{50}	5 4973	1 3877
C_{75}	7 3484	2 1345

be able to answer correctly three quarters of the test questions to be asked about the passage that is being rated, in other words, a score of 100 indicates reading matter that is understandable for persons who have completed fourth grade and are, in the language of the U S Census, barely functionally literate" The range of 100 points was arrived at by multiplying the grade level prediction by 10, so that a point on the formula scale corresponds to one tenth of a grade However, this rela tionship holds true only up to about seventh grade, beyond that, the formula under rates grade level to an increasing degree Finally, the formula—which pre dicted grade level and therefore, difficulty —was turned around by reversing the signs to predict reading ease ' (Before this transformation, the formula read $C_{75} = 0846\ wl + 1015\ sl - 5\ 6835$) The mul tiple correlation coefficient of this formula is $R = 7047$

Formula B (for predicting human in terest") $HI = 3\ 635\ pw + 314\ ps$

Scores computed by this formula, too, have a range from 0 to 100 A score of 100 has the same meaning as in Formula A It indicates reading matter with enough human interest to suit the reading skills and habits of a barely functionally liter- ate' person A score of 0, however, means here simply that the passage contains neither personal words ' nor personal sentences' , in contrast to Formula A, the two elements counted here may be totally absent Since the zero point could be fixed in this way, the scoring was arrived at by dividing the range between 0 (absence of both elements) and 100 (prediction of completed fourth grade) by 100 The for mula therefore contains no statistical con stant The signs were reversed in the same fashion as in Formula A (Before trans formation, this formula read $C_{75} = -1333\ pw - 0115\ ps + 8\ 6673$) The mul tiple correlation coefficient of this formula is $R = 4306$

Since the correlations of three of the four elements with the criterion C_{50} were higher than those with the criterion C_{75}, the multiple correlation with the Criterion C_{50} was computed first As a second step, the values so found were used to predict criterion C_{75}, since it seemed obviously more desirable to predict 75 per cent com prehension than 50 per cent comprehen sion

The correlation between the word length factor (syllable count) and the correspond- ing affix count in the earlier formula was found to be $r = 87$ For practical purposes the two measures may therefore be con sidered equivalent

The number of affixes per 100 words (a) can be predicted from the syllable count (wl) by the formula $a = 6832\ wl - 66\ 6017$ Conversely, the number of syl lables per 100 words (wl) can be pre dicted from the number of affixes (a) by the formula $wl = 1\ 49\ a + 94\ 56$

COMMENT

It is hoped that the two new formulas will prove more useful than the earlier formula Formula A alone, with a correla tion coefficient of 70, has almost as high a prediction value as the combined earlier formula whose correlation coefficient was 74 Formula B has a much lower correla tion coefficient of 43 and, accordingly, does not seem to contribute much to the measurement of readability It should be remembered, however, that because of the criterion used, Formula B predicts only the effect of the two human interest elements on *comprehension* in other words, the cor relation coefficient shows only to what extent human interest in a given text will make the reader understand it better The real value of this formula, however, lies in the fact that human interest will also increase the reader s attention and his motivation for continued reading

In addition, the two new formulas will be more useful for the teaching of writing, since the added factor and the division into two parts will show specific faults in writing more clearly

The significance of Formula A will be more easily understood when it is realized that the measurement of word length is indirectly a measurement of word com plexity (as mentioned above, the correla tion is $r = 87$) and that word complexity in turn is indirectly a measurement of ab straction the correlation between the number of affixes and that of abstract words was found to be 78 (5) Similarly, the measurement of sentence length is in directly a measurement of sentence com plexity In two independent studies the correlation between these two factors was found to be 775 (8) and 72 (15) Sen tence complexity, in turn, may again be considered as a measure of abstraction Formula A, therefore, is essentially a test of the level of abstraction

It seems hardly necessary to prove the importance of human interest in reading, as tested by Formula B That people are most interested in other people is an old truism And the readability value of writ ten dialogue, as tested by the added ele ment, is well described in the following, oddly parallel quotations from a printer and a novelist Have you ever watched people at a library selecting books for home reading? Other things being equal, if they see enough pages that promise interesting dialogue, they are much more apt to put the book under their arm and walk away with it, than if they see too many solid pages which always sug gest hard work' (16) ' What is the use of a book without pictures or conversa tions? thought Alice just before the White Rabbit ran by, in condemnation of the

TABLE 53 4

Comparative Analysis of *The New Yorker* (October 26, 1946) and the *Reader s Digest* (November, 1946)

	New Yorker	*Reader's Digest*
Old Formula		
Average sentence length in words	20	16
Affixes per 100 words	36	34
Personal words per 100 words	10	8
Readability score	3 59	3 05
New Formula A		
Average sentence length in words	20	16
Syllables per 100 words	148	145
Reading ease score	61	68
New Formula B		
Personal words per 100 words	10	8
Personal sentences per 100 sentences	39	15
Human interest' score	49	34

book her sister was reading, and this child ish comment is supported by novel readers of all degrees of intelligence Long close paragraphs of print are in themselves apt to dismay the less serious readers and their instinct here is a sound one, for an excess of summary and an insufficiency of scene in a novel make the story seem remote, without bite second hand A great part of the vigor, the vivacity and the readability of Dickens derives from his in numerable interweavings of scene and sum mary, his general method is to keep summary to the barest essential minimum, a mere sentence or two here and there be tween the incredibly fertile burgeoning of his scenes' (2)

In preliminary tests of the formulas, the following results were found

When the newly isolated fourth element (personal sentences) was applied to the psychology texts by Koffka and James men tioned above (17), it was found that the percentage of 'personal sentences in Koffka was negligible (4 per cent), whereas in James s first volume it was 16 per cent and in his second volume 10 per cent A striking example of this difference in style is the following of James s per sonal sentences Ask half the common drunkards you know why it is that they fall so often prey to temptation, and they will say that most of the time they cannot tell This sentence shows well the aspect of readability that eluded the earlier for mula

When the old and the new formulas were applied to two random copies of the *New Yorker* (October 26, 1946) and the *Reader's Digest* (November 1946), the re sults were as shown in Table 53 4

As can be seen, the old formula rated the *Reader s Digest* significantly more read able than the *New Yorker* the new for mula A also shows that the *Reader's Digest* is significantly easier to read But the new formula B clearly shows a large difference in human interest in favor of the *New Yorker*

The Formulas Restated

For practical application the formulas may be restated this way

To measure the readability (reading ease and human interest) of a piece of writing, go through the following steps

Step 1 Unless you want to test a whole piece of writing, take samples Take enough samples to make a fair test (say, 3 to 5 of an article and 25 to 30 of a book) Don t try to pick good or typ ical samples Go by a strictly numerical scheme For instance, take every third paragraph or every other page Each sample should start at the beginning of a paragraph

Step 2 Count the words in your piece of writing or, if you are using samples take each sample and count each word in it up to 100 Count contractions and hyphenated words as one word Count as words num bers or letters separated by space

Step 3 Count the syllables in your 100 word samples or, if you are testing a whole piece of writing, compute the number of syllables per 100 words If in doubt about syllabication rules, use any good dictionary Count the number of syllables in symbols and figures according to the way they are normally read aloud e g , two for $ (dol lars') and four for 1918 (nineteen eight een) If a passage contains several or lengthy figures, your estimate will be more accurate if you don t include these figures in your syllable count In a 100 word sample be sure to add instead a corre sponding number of words in your syllable count To save time, count all syllables except the first in all words of more than one syllable and add the total to the num ber of words tested It is also helpful to 'read silently aloud' while counting

Step 4 Figure the average sentence length in words for your piece of writing or, if you are using samples, for all your samples combined In a 100 word sample, find the sentence that ends nearest to the 100 word mark—that might be at the 94th word or the 109th word Count the sen tences up to that point and divide the number of words in those sentences by the number of sentences In counting sentences, follow the units of thought rather than the punctuation usually sentences are marked off by periods, but sometimes they are marked off by colons or semicolons—like

these But don't break up sentences that are joined by conjunctions like *and* or *but*

Step 5 Figure the number of personal words per 100 words in your piece of writing or, if you are using samples, in all your samples combined Personal words" are (a) All first-, second , and third person pronouns except the neuter pronouns *it its itself* and *they them their, theirs, themselves* if referring to things rather than people (b) All words that have masculine or feminine natural gender, *e g , Jones, Mary father, sister, iceman actress* Do not count common gender words like *teacher doctor employee assistant spouse* Count singular and plural forms (c) The group words *people* (with the plural verb) and *folks*

Step 6 Figure the number of 'personal sentences per 100 sentences in your piece of writing or, if you use samples, in all your samples combined Personal sentences are (a) Spoken sentences, marked by quotation marks or otherwise, often including so called speech tags like he said (e g , I doubt it —We told him 'You can take it or leave it"— That's all very well, he replied, showing clearly that he didn t believe a word of what we said) (b) Questions, commands, requests, and other sentences directly addressed to the reader (c) Exclamations (d) Grammatically incomplete sentences whose full meaning has to be inferred from the context (e g , Doesn t know a word of English —Handsome, though —Well, he wasn t —The minute you walked out) If a sentence fits two or more of these definitions, count it only once Divide the number of these 'personal sentences' by the total number of sentences you found in Step 4

Step 7 Find your 'reading ease score by inserting the number of syllables per 100 words (word length, *wl*) and the average sentence length (*sl*) in the following formula

R E ("reading ease') = 206 835 − 846 *wl* − 1 015 *sl*

The "reading ease" score will put your piece of writing on a scale between 0 (practically unreadable) and 100 (easy for any literate person)

Step 8 Find your "human interest" score by inserting the percentage of personal words (*pw*) and the percentage of personal sentences' (*ps*) in the following formula

H I (' human interest') = 3 635 *pw* + 314 *ps*

The 'human interest' score will put your piece of writing on a scale between 0 (no human interest) and 100 (full of human interest)

In applying the formulas, remember that Formula A measures *length* (the longer the words and sentences, the harder to read) and Formula B measures *percentages* (the more personal words and sentences, the more human interest)

Roughly, reading ease' scores will tend to follow the pattern shown in Table 53 5 Human interest" scores will follow the general pattern shown in Table 53 6

Sample Application

As an example of the application of the

TABLE 53 5

Pattern of Reading Ease' Scores

Reading Ease' Score	*Description of Style*	*Typical Magazine*	*Syllables per 100 Words*	*Average Sentence Length in Words*
0 to 30	Very difficult	Scientific	192 or more	29 or more
30 to 50	Difficult	Academic	167	25
50 to 60	Fairly difficult	Quality	155	21
60 to 70	Standard	Digests	147	17
70 to 80	Fairly easy	Slick fiction	139	14
80 to 90	Easy	Pulp fiction	131	11
90 to 100	Very easy	Comics	123 or less	8 or less

TABLE 53 6

Pattern of Human Interest Scores

Human Interest Score	*Description of Style*	*Typical Magazine*	*Percentage of Personal Words*	*Percentage of Personal Sentences*
0 to 10	Dull	Scientific	2 or less	0
10 to 20	Mildly interesting	Trade	4	5
20 to 40	Interesting	Digests	7	15
40 to 60	Highly interesting	*New Yorker*	11	32
60 to 100	Dramatic	Fiction	17 or more	58 or more

new formulas, two recent descriptions of the nerve block method of anesthesia will be used By an odd coincidence these two variations upon a theme appeared within the same week in *Life* (October 27, 1947) and *The New Yorker* (October 25, 1947) The *Life* story served as text accompanying a series of pictures, it is straight reporting not particularly simple, and lacks human interest (which was supplied by the pictures) The *New Yorker* passage is part of a personality profile, vivid, dramatic, using all the tricks of the trade to get the reader interested and keep him in suspense

From *Life*

Except in the field of surgery, control of pain is still very much in the primitive stages Countless thousands of patients suffer the tortures of cancer, angina pectoris and other distressing diseases while their physicians are helpless to relieve them A big step toward help for these sufferers is now being made with a treatment known as nerve blocking This treatment, which consists of putting a block' between the source of pain and the brain, is not a new therapy But its potentialities are just now being realized Using better drugs and a wider knowledge of the mechanics of pain gained during and since the war, Doctors E A Rovenstine and E M Papper of the New York University College of Medicine have been able to help two thirds of the patients accepted for treatment in their pain clinic at Bellevue Hospital

The nerve block treatment is comparatively simple and does not have serious aftereffects It merely involves the injection of an anesthetic drug along the path of the nerve carrying pain impulses from the diseased or injured tissue to the brain Although its action is similar to that of spinal anesthesia used in surgery, nerve block generally lasts much longer and is only occasionally used for operations The N Y U doctors have found it effective in a wide range of diseases, including angina pectoris, sciatica, shingles, neuralgia and some forms of cancer Relief is not always permanent, but usually the injection can be repeated Some angina pectoris patients have had relief for periods ranging from six months to two years While recognizing that nerve block is no panacea, the doctors feel that results obtained in cases like that of Mike Ostroich (*next page*) will mean a much wider application in the near future

From *The New Yorker*

Recently, [Rovenstine] devoted a few minutes to relieving a free patient in Bellevue of a pain in an arm that had been cut off several years before The victim of this phantom pain said that the tendons ached and that his fingers were clenched so hard he could feel his nails digging into his palm Dr Rovenstine's assistant, Dr E M Papper, reminded Rovenstine that a hundred and fifty years ago the cure would have been to dig up the man s arm if its burial place was known, and straighten out the hand Rovenstine smiled ' I tell you, he said *We ll* use a two percent solution of procaine, and if it works, in a couple of weeks we ll go on with an alcohol solution Procaine, you know, lasts a couple of weeks, alcohol six months or longer In most cases of this sort, I use the nerve block originated by

Labat around 1910 and improved on in New Orleans about ten years back, plus one or two improvisations of my own (Nerve blocking is a method of anesthetizing a nerve that is transmitting pain)

The man with the pain in the non existent hand was an indigent, and Rovenstine was working before a large gallery of student anesthetists and visitors when he exorcised the ghosts that were paining him Some of the spectators, though they felt awed, also felt inclined to giggle Even trained anesthetists sometimes get into this state during nerve block demonstrations because of the tenseness such feats of magic induce in them The patient, thin, stark naked, and an obvious product of poverty and cheap gin mills, was nervous and rather apologetic when he was brought into the operating theatre He lay face down on the operating table Rovenstine has an easy manner with patients, and as his thick, stubby hands roamed over the man s back he gently asked, How you doing? My hand, it is all closed together Doc,' the man answered, startled and evidently a little proud of the attention he was getting You ll be O K soon, Rovenstine said, and turned to the audience 'One of my greatest contributions to medical science has been the use of the eyebrow pencil, he said He took one from the pocket of his white smock and made a series of marks on the patient s back, near the shoulder of the amputated arm, so that the spectators could see exactly where he was going to work With a syringe and needle, he raised four small weals on the man's back and then shoved long needles into the weals The man shuddered but said he felt no pain Rovenstine then attached a syringe to the first needle, injected the procaine solution, unfastened the syringe attached it to the next needle, injected more of the solution, and so on The patient s face began to relax a little Lord, Doc, he said My hand is loosening up a bit already You ll be all right by tonight, I think, Rovenstine said He was

A comparative analysis of these two passages is shown in Table 53 7

The two passages furnish a good illustration of the stylistic features measured and emphasized by the two new formulas

REFERENCES

1 Alden, J Lots of Names—Short Sentences—Simple Words, *Printer s Ink* June 29, 1945, 21–22

2 Bentley, Phyllis *Some Observations on the Art of the Narrative* New York The Macmillan Company, 1947

3 Cowing Amy G , They Speak his Language, *Journal of Home Economics* 1945, Vol 37, 487–489

TABLE 53 7

Comparative Analysis of Treatment of Same Theme in *Life* and *The New Yorker*

	Life (290 words)	*New Yorker* (495 words)
Old Formula		
Average sentence length in words	22	18
Affixes per 100 words	48	35
Personal words per 100 words	2	11
Readability score	5 16	3 20
New Formula A		
Average sentence length in words	22	18
Syllables per 100 words	165	145
'Reading ease score	46	66
New Formula B		
Personal words per 100 words	2	11
Personal sentences per 100 sentences	0	41
Human interest' score	7	53

4 Fihe, Pauline J, Wallace Viola, and Schulz, Martha, compilers *Books for Adult Beginners Grades I to VII* Rev ed Chicago American Library Association 1946
5 Flesch, R *Marks of Readable Style a Study in Adult Education* New York Bureau of Publications Teachers College, Columbia University, 1943 (Contr to Educ No 897)
6 Flesch, R *The Art of Plain Talk* New York Harper & Brothers, 1946
7 Flesch R How to Write Copy that will be Read, *Advertising & Selling* March 1947, 113ff
8 Gray W S and Leary, Bernice E *What Makes a Book Readable* Chicago University of Chicago Press, 1935
9 Gunning, R 'Gunning Finds Papers Too Hard to Read *Editor & Publisher* May 19, 1945 12
10 Lorge, I Predicting Reading Difficulty of Selections for Children, *Elementary English Review* 1939, Vol 16 229–233
11 McCall W A and Crabbs, Lelah M *Standard Test Lessons in Reading* Books II III IV, and V New York Bureau of Publications Teachers College, Columbia University, 1926
12 Miller L R Reading Grade Placement of the First 23 Books Awarded the John Newbery Prize, *Elementary School Journal* 1946, 394–399
13 Murphy, D R Test Proves Short Sentences and Words Get Best Readership, *Printer s Ink* 1947, Vol 218, 61–64
14 Murphy, D R How Plain Talk Increases Readership 45% to 66% *Printer s Ink* 1947 Vol 220, 35–37
15 Sanford F H *Individual Differences in the Mode of Verbal Expression* Unpublished Ph D thesis Harvard University 1941
16 Sherbow B *Making Type Work* New York Century, 1916
17 Stevens, S S, and Stone Geraldine Psychological Writing, Easy and Hard *American Psychologist* 1947, Vol 2 230–235 Discussion, 1947, Vol 2, 523–525
18 Foreign News Written Over Heads of Readers *Editor & Publisher* Dec 28, 1946, Vol 28
19 *How Does Your Writing Read?* U S Civil Service Commission Washington U S Government Printing Office, 1946
20 *Readability in News Writing Report on an Experiment by United Press* New York United Press Associations, 1945

Reliability of the Flesch Readability Formulas *

PATRICIA M HAYES, JAMES J JENKINS, and BRADLEY J WALKER

Hayes and Jenkins are primarily responsible for the first study in this paper and Walker is primarily responsible for the second

A formula developed by Flesch (3) for estimating the comprehension difficulty of written material has received widespread attention in many areas of communication It has been applied in the fields of journalism (7), advertising (1), industrial communications (5), government publications (8) and many others

In view of the wide use of his formula, Flesch (4) published a revision in 1948 designed to increase its utility and make its application and interpretation easier This revision proposes the use of two formulas to measure two relatively independent aspects of readability The first formula involves word length (*wl*) and sentence length (*sl*) and gives a measure of "reading ease' (RE) The second formula, based on personal words (*pw*) and personal sentences (*ps*), yields a measure of human interest' (HI)

As these formulas become increasingly

* Reprinted from *Journal of Applied Psychology* Vol 34, No 1, February, 1950

popular, they must, of course, be evaluated critically Like other psychological tools, they must be tested for validity and reliability Flesch (4) reports several studies of the validity of the original formula which indicate that material rated more readable by the formula also proves more readable in terms of readership surveys and opinions of judges As yet, however, no studies of the reliability of the formulas as applied by different analysts have been reported

The following studies were designed as first steps in the examination of analyst-to analyst reliability of the formulas to determine the extent to which they are effectively objective

First Study

The material chosen for analysis in the first study consisted of the 40 prize winning letters in the recent General Motors' Why I Like My Job' contest (9) These letters were selected because they presented a wide range of difficulty, style, structure and content It was believed the letters would afford a maximum number of problems in interpretation and would provide a rigid test of objectivity

Two sets of samples were drawn from the 40 letters Each set consisted of two 100 word samples from each letter Since the letters ranged in length from about 350 to 3000 words with a median length of 750 words, there was little overlapping between the sets of samples

Two experienced and two inexperienced analysts participated in the study[1] The experienced workers analyzed both sets of samples, the inexperienced each analyzed one set The experienced analysts had worked with the original formula and the revised formulas for a year and a half The inexperienced analysts had never worked with the formulas before The analysts made no attempt to agree on interpretation of Flesch s instructions and refrained from discussing interpretation with anyone else

[1] The writers would like to acknowledge the assistance of Barbara Lee and James Farr in this part of the project

Reading ease and human interest scores were computed from tables (2) in an effort to minimize computational errors Results of analyses made by different investigators on the same set of samples were compared by determining the significance of differences between means of the four variables (*wl sl pw ps*) and the two scores (HI, RE) for the set of samples as a whole, the extent of correlation between results of analysts, the number and degree of differences in actual scores for each sample, and the number of differences in descriptive categories assigned to each sample

Results of First Study

The assumption of wide variability in the material used was confirmed As may be seen in Table 54 1, samples ranged from 6 to 88 on the reading ease scale of 0 to 100 points and from 13 to 100 on the human interest scale of 0 to 100 points

The means of the four variables and two scores obtained by analysts on the same sets of samples were tested for significant differences by use of the critical ratio corrected for correlation None of the differences between any analysts in either sample set proved to be significant at the 5 per cent level

Rank difference correlations were computed between each pair of analysts within each sample set on the rank given each letter These correlation coefficients are presented in Table 54 2 All of the correlations are positive and significantly different from zero beyond the one per cent level

An inspection of Table 54 2 indicates that analysts were in good agreement in interpreting the components and final score for reading ease For the human interest variables, however, there was much less agreement between analysts Personal words were apparently interpreted much the same by analysts, but it is evident there were diverse interpretations of personal sentences This acts, of course, to lower the correlations of the human interest score

It should be noted that correlations between experienced analysts (B and C) are not appreciably different from those with inexperienced analysts (A and D)

TABLE 54.1

Means, Standard Deviations and Ranges for the Four Variables and Two Scores of the Flesch Readability Formulas for Each Analyst

Mean	*Ana-lyst*	*wl*	*sl*	*RE*	*pw*	*ps*	*HI*
Set 1	A	145.0	23.7	60.2	9.9	15.3	41.1
	B	145.2	23.3	60.4	9.7	10.2	38.2
	C	145.0	23.9	59.9	9.5	13.6	39.5
Set 2	B	144.0	22.7	62.1	9.5	11.8	38.6
	C	143.8	22.5	62.3	10.2	14.2	42.2
	D	144.3	23.2	61.2	9.5	13.1	39.3
Standard deviation							
Set 1	A	11.9	11.1	15.9	3.5	26.6	18.2
	B	11.8	11.3	16.1	3.3	17.6	16.4
	C	11.7	11.1	15.4	3.7	18.4	15.6
Set 2	B	9.4	10.8	13.6	2.9	21.4	14.3
	C	9.0	9.8	12.9	2.8	26.7	13.6
	D	9.7	9.9	13.4	3.4	21.1	14.0
Range							
Set 1	A	124–174	11–72	7–85	4–20	0–100	15–100
	B	125–174	10–72	6–84	4–20	0– 83	13– 97
	C	124–173	11–72	6–82	5–18	0– 91	18– 94
Set 2	B	122–164	12–80	42–81	5–15	0–100	16– 79
	C	122–162	11–74	42–88	5–17	0–100	18– 80
	D	123–164	12–74	42–86	5–15	0–100	16– 80

Since neither of the statistical methods presented above reveals the actual differences between analysts for a given sample, a third kind of comparison was made. Within each sample set all analysts were compared on reading ease and human interest scores for each letter. Actual point differences for each pair of analysts were tabulated. Results of the 240 comparisons are shown in Table 54.3.

This Table also suggests that there is greater agreement on reading ease (90 per cent of the comparisons within four points of each other) than on human interest (90 per cent of the comparisons within 8 points of each other). On the 100-point scale designed to be used as an estimating device, deviations as small as these do not appear to be of great importance.

Again it should be mentioned that no consistent difference was found in the number or extent of deviations between scores of experienced analysts compared to

TABLE 54.2

Rank Order Correlation Coefficients between Pairs of Analysts for the Variables and Scores of the Flesch Readability Formulas *

	Sample Set 1			*Sample Set 2*		
*Ana-lysts***	A and B	A and C	B and C	C and D	B and D	B and C
wl	.99	.99	.99	.99	.99	.99
sl	.94	.97	.94	.83	.87	.94
RE	.98	.99	.98	.93	.95	.97
pw	.93	.94	.96	.92	.99	.93
ps	.89	.69	.74	.63	.87	.60
HI	.88	.91	.96	.78	.97	.80

* These correlations should be interpreted with caution since the data are markedly skewed in the case of *ps* and HI. It should be noted, however, that the order of relative accuracy for the scores is the same whether correlations, point differences or category differences are considered.

** Analysts A and D are inexperienced; B and C, experienced.

their deviations with scores of inexperienced analysts

A final comparison was made between analysts in terms of descriptive categories in which results are often reported and utilized Flesch (4) divides the reading ease range into seven levels varying from very difficult" to very easy" and the human interest range into five levels varying from dull' to dramatic'

TABLE 54 3

Differences in Score Points Between Analysts on Identical Samples *

	Reading Ease		*Human Interest*	
Difference in Points	Per Cent of Comparisons	Cumulative Percentage	Per Cent of Comparisons	Cumulative Percentage
1	49 2	49 2	33 3	33 3
2	25 8	75 0	18 0	51 3
3	9 2	84 2	6 6	57 9
4	6 2	90 4	15 4	73 3
5	3 8	94 2	4 2	77 5
6	1 6	95 8	5 4	82 9
7	1 7	97 5	3 4	86 3
8	8	98 3	3 7	90 0
9	9	99 2	2 5	92 5
10	8	100 0	8	93 3
11 or more			6 7	100 0

* Based on 240 comparisons three analyses for each of 80 samples of 200 words

Of the 240 comparisons, in only 14 cases (5 8 per cent) did the analysts differ in the category assigned to reading ease In 28 cases (11 7 per cent) they differed in the category assigned to human interest Of these cases of disagreement, only 2 (8 per cent) were greater than one category for reading ease, and only 4 (1 7 per cent) were greater than one category for human interest

Discussion

While the results of this first study seem to constitute a limited but fairly clear answer to the question of reliability of the Flesch formulas, mention of the greatest sources of error may be of some value in interpreting the data and may provide a few hints to those who wish to use the formulas

The greatest discrepancies obviously appear in interpretation of personal sentences A study of Table 54 2 shows that correlations for this variable are especially low when analyst "C' is involved Samples which contributed most to the discrepancy between C" and the other analysts were studied Over half of the major differences between "C' and the others involved one type of personal sentence defined by Flesch (4) as 'grammatically incomplete sentences whose full meaning has to be inferred from the context The examples given by Flesch (4) appear to be taken from conversations and apparently the definition was regarded as limited to conversations by analysts "A," 'B' and 'D" It would seem, however, that the examples are to this extent misleading since conversational sentences are already covered by Flesch s first definition regarding spoken sentences It might be suggested that analysts study the definitions carefully and that Flesch provide more varied examples

A second source of disagreement involved rhetorical questions Analyst 'C" did not count these as personal sentences, but it appears clear from Flesch s definition that these should have been considered and scored as the other analysts scored them

If these two sources of error (incomplete sentences and rhetorical questions) had been corrected, correlations for personal sentences would have been raised above 90, and the human interest scores would have appeared much more reliable

Errors in personal words were few and appear to be due largely to carelessness While there were no consistent errors, from time to time one analyst or another tended to regard a common gender noun like worker or 'manager' as a personal word Flesch's definitions and examples are explicit on this point (4)

Errors in sentence length resulted chiefly from disregarding directions on counting the last sentence when the sample ends in the middle of a sentence and from disagreement on breaking sentences into units of thought It may be noted that the lowest correlations in Table 54 2 for sentence length involve analyst 'D" A study of the

samples yielding the greatest discrepancies revealed that if analyst 'D had broken sentences into units of thought in just two instances, correlations would have been above 95 Here it appears that more careful attention to directions would have assured high reliability

Errors in word length are all very small and appear to reflect minor clerical errors

Second Study

A second study was conducted to test our findings with a large number of inexperienced analysts Samples of 500 words from 63 house organs and employee publications which were being examined in connection with a continuing study of industrial communication (6) were assigned for analysis to 18 members of a graduate seminar in psychology Each student analyzed seven publications which were subsequently reanalyzed by another member of the seminar Assignments were anonymous and cooperation between students was discouraged Only three of the students had appreciable experience with the formulas prior to the time of the study

TABLE 54 4

Sampling Statistics of Test and Re Test Distributions for the Second Study

	Means		*Standard Deviation*		*Range*	
	Test	Re Test	Test	Re Test	Test	Re Test
el	155 4	154 9	7 98	7 98	138–17-	140–167
sl	20 6	20 5	4 39	4 15	15–45	13–45
RE	54 5	55 2	8 40	7 71	30–73	31–69
pw	7 3	6 6	2 38	2 24	3–16	-–13
ps	12 8	13 1	11 17	11 45	0–46	0–48
HI	30 3	28 3	10 74	9 56	15–64	8–62

The results of the analyses provided pairs of scores for each of the publications The first analyses were compared with the second analyses to determine the reliability of the application of the formulas to the same samples Product moment correlations between the 'test and 'retest" analyses are as follows *wl* 90, *sl*, 92, RE, 91, *pw* 78, *ps* 64, and HI, 81 All coefficients were positive and significantly different from zero

The data for the means, standard deviations and ranges are presented in Table 54 4 A comparison of the standard deviations and the ranges in Table 54 4 with those in Table 54 1 reveals that the material used in the second study was appreciably more homogeneous than that used in the first The correlations found in the second study, then, might be expected to be smaller than those of the first study Accordingly, the correlations presented immediately above were corrected by estimating their magnitude on the basis of the more heterogeneous material of the first study This correction gave the following coefficients *wl*, 95, *sl* 99 RE, 98, *pw* 88, *ps* 85, and HI, 92

These correlations approximate those found in the first study and would lead one to the same conclusions Reading ease with its components is analyzed quite reliably and human interest with its components is analyzed with less, though still fair reliability Analysis of personal sentences again shows the greatest lack of agreement between analysts

The difference in points between 'test and 'retest' analyses agrees rather closely with the data from the first study given in Table 54 3 Ninety per cent of the paired scores for reading ease were within 6 points of each other and approximately 90 per cent of the paired scores for human interest were within 8 points of each other

Summary and Conclusions

An examination of analyst-to-analyst reliability of the Flesch readability formulas was conducted In the first study two sets of samples were drawn from reading material of a highly variable nature believed to involve a large number of problems of interpretation The sets of samples were analyzed by two inexperienced and two experienced analysts Results of analysts for each set of samples were compared by testing the significance of mean differences on the variables and the scores, correlating results on the variables and scores, tabulating deviations in terms of score points and tabulating disagreements in descriptive categories

In the second study, 18 students analyzed

samples of 500 words from 63 industrial house organs Each sample was independ ently analyzed by two analysts Correla tions between the first and second sets of analyses were computed and then cor rected for restriction of range Deviations in terms of score points were computed

From the above data the following conclusions seem justified

1 Analyst to analyst reliability on word length, sentence length, and reading ease is quite high for the kinds of material used in this study

2 Analyst to analyst reliability on personal words is fair, but on personal sen tences (and as a result of human interest) is lower than might ordinarily be consid ered desirable

3 For practical purposes the Flesch for mulas and the directions for their use are sufficiently objective to be used even by inexperienced analysts to obtain estimates of the reading ease and human interest of written material

REFERENCES

1 Alden, J, Lots of Names—Short Sentences—Simple Words, *Printer's Ink* June 29, 1945, 21–22

2 Farr, James N, and Jenkins, James J, 'Tables for Use with the Flesch Readability Formulas, *Journal of Applied Psychology* 1949, Vol 33, 275–278

3 Flesch, R, *The Art of Plain Talk* New York Harper and Brothers 1946

4 Flesch, R, A New Readability Yardstick,' *Journal of Applied Psychology* 1948, Vol 32, 221–233

5 Paterson, D G, and Jenkins, James J, 'Communication Between Management and Workers ' *Journal of Applied Psychology* 1948, Vol 32, 71–80

6 Paterson, D G, and Walker, B J, Readability and Human Interest of House Organs, *Personnel* 1949, Vol 25 438–441

7 Swanson, C E, Readability and Readership a Controlled Experiment *Journalism Quarterly* 1948, Vol 25, 339–343

8 *How Does Your Writing Read?* U S Civil Service Commission Washington U S Government Printing Office, 1946

9 *The Worker Speaks* General Motors, 1947

How Readable Are Corporate Annual Reports? *

SIROON PASHALIAN and WILLIAM J E CRISSY

This paper is a report on some of the findings from Pashalian s M A thesis entitled, 'An Investigation of the Application of the New Flesch Readability Formulas to Corporate Annual Reports,' submitted to the Department of Psychology, Graduate School of Arts and Science at New York University, October 1949 Explorative findings concerned with the relationship between Flesch indices and judges ratings of readability have not been reported due to basic design limitations

Business enterprises have long been con cerned with communication problems To day there is increasing interest in how best to "get the word around to jobholders, shareholders, customers, and the general public One of the first formal media of communication to be used was the cor porate annual report In recent years an extensive literature has developed con cerning how best to construct and publish such reports (1, 2, 7, 8) Most of the papers have reflected judgments of varying degrees of expertness rather than findings based upon experimental research

A current problem in the construction of the annual report is that of endeavoring to make it more understandable and more widely read Readership surveys show a

* Reprinted from *Journal of Applied Psychology* Vol 34, No 4, August 1950

considerable apathy to company reports Yet, it is by no means an easy task to prepare adequate and concise reports The challenge is one of presenting sufficient technical data and information for the financial expert, satisfying the requirements of the law (especially in railroad and public utility reporting), and at the same time, being meaningful to those whose interests are of a more general nature The present demand crystallizes as the need for writing an informative account of the year s operations, and for presenting the material in such a way that the report will be read

In connection with this problem, one of the writers undertook to investigate the readability of corporate annual reports by means of the new "Flesch Readability Formulas (5, 5a)

Method

The annual reports of those corporations that are listed in the Corporate Billion Dollar Club in the June 11, 1949 issue of *Business Week* were included in the present study These members are either non financial companies with assets of over $1 billion, or with annual revenues or sales of over $1 billion, or both Presumably then, these corporations have the largest number of stockholders, employees and other persons interested in their operations In other words, there is substantial public interest in these big corporations, their annual reports are expected to reach a vast audience

Applying the sampling technique suggested by Flesch (3, 4, 5), one hundred word samples were chosen from every other page of each of these 26 reports This procedure, with no restrictions on the number of samples to be taken per report, is believed to have achieved a fair sampling in proportion to the length and breadth of the report A total of 211 samples were examined, the average number of samples per report was 8 1

Results

The findings on the application of the new Flesch Readability Formulas to the 26 annual reports are listed in Table 55 1

Analysis of the Reading Ease Measures

The range of readability scores was from 6 to 58 According to Flesch reference categories for these scores (5, 5a), these reports vary within descriptive styles of *very difficult* as with material of scientific and professional journals to *fairly difficult* as in literary and quality magazines, such as *Harper s* This range interpreted in terms of the educational attainments of the U S adult population suggests a potential audience of from 4½ per cent of the population completing college, to 40 per cent of the population who have had some high school education (4)

The average Reading Ease score for the entire set of reports was 34 37 Writing at this level is generally *difficult* and descriptive of the style in academic material, for example, the *Yale Review* which may be comprehended by 24 per cent of the population who have graduated from high school or have had some college training

The measure of average sentence length ranged from 16—*fairly easy* typical of slick fiction and understandable by 80 per cent of the population, to 53—*very difficult* above much scientific material and understood by approximately 4½ per cent of the population

The measure of the average number of syllables per 100 words ranged from 156—*fairly difficult* to 183—*difficult* Flesch has advised that a comfortable text contains one and one half times as many syllables as words (6) The preponderance of reports in the *difficult* category on this measure (22 of the 26 reports) is indicative of a high level of abstraction in the language of these reports

In relation to this syllable measure, the factor of numbers was encountered in the material of the annual reports Under Flesch s directions, numbers separated by space are counted as words in any text, several and lengthy figures should be omitted from the syllable count Instead, a corresponding number of words to the number of figures omitted should be added, and their syllable totals added to those already counted (5) When applying these directions to the samples, the number of

TABLE 55 1

Summary of Reading Ease and Human Interest Measures on the Twenty six Annual Reports

Industry and Average R E	*Corporate Annual Report*	*Average Sentence Length*	*Average Number Syllables*	*Reading Ease Score*	*Percentage of Personal Words*	*Sent*	*Human Interest Score*
Merchandise	Sears, Roebuck	22	162	47	2	0	7
44 50	Montgomery Ward	16	175	43	0	0	0
Communications 43 00	Bell Telephone	24	165	43	1	0	4
Foods	Swift & Co	17	156	58	5	2	19
43 00	Armour & Co	21	168	43	4	0	15
	Safeway Stores	27	182	28	2	0	7
Autos & Accessories	General Motors	21	174	38	2	0	7
40 00	Chrysler Corp	20	171	42	1	0	4
Oil	Standard Oil (N J)	21	179	34	0	0	0
33 50	Standard Oil (Ind)	25	173	35	2	0	7
	Socony Vacuum	21	180	33	2	0	7
	Texas Co	23	182	32	1	0	4
	Gulf Oil	24	175	34	1	0	4
	Standard Oil (Cal)	24	178	32	0	0	0
Utilities	Consolidated Edison	27	166	39	1	0	4
31 50	Commonwealth Southern	31	183	24	0	0	0
Railroads	Pennsylvania	27	170	36	1	0	4
28 50	NY Central	25	174	34	1	0	4
	Southern Pacific	32	175	26	0	0	0
	Santa Fe	32	177	25	1	0	4
	Baltimore & Ohio	18	171	44	2	0	7
	Union Pacific	53	174	6	0	0	0
Machinery & Supplies 26 00	General Electric	30	178	26	0	0	0
Metals & Chemicals	U S Steel	32	173	28	0	0	0
25 67	E I du Pont	31	176	26	1	0	4
	Bethlehem Steel	39	172	23	1	0	4

figures per 100 words was also recorded The average number of figures per 100 words ranged from 2 30 to 8 80 The high est number of figures per 100 words ranged from 5 to 21 Thus, a surprisingly large number of figures appeared in small samples of 100 words Their disastrous effect on the general reader not widely trained in numbers or mathematics invites speculation It would seem, therefore, that greater care and attention should be given to determining best ways of presenting fig ures in such reports Hundred word sam ples that are crowded with 10 to 20 lengthy figures should caution writers or editors, and suggest their more effective incorpo ration in a table or chart

The significance of these results on read-ability obtained by analysis utilizing the Flesch technique is perhaps best indicated by showing what would improve the scores made A need is demonstrated for more effective use of punctuational de vices The semicolon, for instance, can be more widely used to shorten sentence lengths and at the same time, to maintain any indications that the words and infor mation belong closely together

Similarly, the need for writing at a less difficult level is indicated A writing down would not necessarily be an under estimation of intelligence People whom corporations want to influence probably range from low normal intelligence to the superior They have the capacity to grasp such concepts as gross sales, profits, etc However, the vast audience which these reports reach is assumed by corporate re porters to possess far greater language fa cility than it does The language of these reports, shown in terms of the education of the U S adult population above, is too

difficult for the great bulk of this diversified readership to comprehend Corporate writers are overestimating the language experience of their potential audience—stockholders, employees and the general public

Human Interest Values

The range of Human Interest scores was from 0 to 19 Again, in terms of Flesch references (5, 5a), these styles are from *dull* descriptive of the style in scientific journals, to *mildly interesting,* descriptive of trade magazines The average Human Interest score for the entire set of reports was 4 27—*dull*

Thus, the corporate writing of these twenty six annual reports is extremely low in human interest value, i e, in the personal words and sentences which provoke and continue general reader interest, and help the reader to understand the text better In an era which is fostering the teamwork and cooperation of stockholders and employees alike, the need for the stress on "we" and "our" and "us" is inescapable "I" can serve to bridge the tremendous gap between the President and stockholders and employees This set of reports used such words sparsely Personal words help to convey the feeling in the material of having been written directly to the reader, whoever he may be They can reflect the whole spirit and tone of the organization

In addition, corporate writing can direct greater attention to individual personalities Although the sampling necessarily tapped only certain pages, extremely few samples mentioned specific persons and their accomplishments It appeared that much of this kind of information was confined to Employee Relations headings or obituaries People are interested in people, and they want to become better acquainted with the outstanding personalities of the corporation Yet, among the 21,100 words sampled in this study, only approximately 20 names were mentioned, and even these were noticeably concentrated within certain reports

Moreover, to enhance human interest value, there remains the need for the appropriate use of personal sentences—exclamations, questions and commands directly expressed to the reader Only one report among the 26 possessed a sentence of this description in the sampling scheme—a question More question marks in annual reports can provoke thought, continue or revive reader interest Similarly, direct commands are another interest controlling device Instead of the impersonal, It will be noted , Note can do much more to invoke the effort of a glance at the charts and an independent analysis

Sample Passages

To illustrate the various levels of difficulty obtained by means of the new Flesch Readability Formulas, sample passages from the annual reports of those corporations which ranked twenty sixth, thirteenth, and first in Reading Ease are furnished below

From the Union Pacific Railroad report

Capital Stock

At the annual meeting of Union Pacific Railroad Company stockholders held on May 11, 1948, the Articles of Association were amended so that on July 1 1948 the total number of authorized shares of preferred and common stocks of the Company were doubled (with no increase in the total aggregate par value thereof), and the then outstanding 995,431 shares of $100 par value preferred stock became 1,990,862 shares of $50 par value preferred stock, and the then outstanding 2,222,910 shares of $100 par value common stock became 4,445,820 shares of $50 par value common stock, each of the new $50 par value preferred and common shares being entitled to one vote at any meeting of stockholders

From the New York Central Railroad report

Dieselization is progressing

Carrying forward our motive power modernization, the Central and leased lines together with two affiliates, the Pitts

burgh & Lake Erie and the Indiana Harbor Belt Railroads, ordered in 1948 new Diesel-electric locomotives at a total cost of approximately $33,600,000. The bulk of these locomotives, on which deliveries will extend into 1950, are for road freight and for switching service. The Central's portion was about $24,790,000.

Locomotives delivered during 1948 increased the Dieselized portion of the total road freight train mileage of the Central and leased lines to approximately 13.5 per cent by the end of the year.

From the Swift and Company report:

What Swift & Co. is Trying To Do

The public rightly expects a business to accomplish certain desirable things.

Who determines what is desirable in a free country? Not one man or a group of men. Each individual decides for himself whether he will buy from, sell to, work for, or invest in a company.

A decision to buy a product is a vote in its favor. The votes of millions of people may cause prices to go up or down. The results quickly tell what the public thinks desirable.

Such economic democracy can thrive only in a certain climate—one in which prices are free and competitive and business is spurred by the hope of profits and the fear of losses.

Analysis by Industries

The arrangement of the twenty-six reports by industry in Table 55.1 facilitates interesting and noteworthy comparisons. There seems to be a certain amount of homogeneity within industries on all the obtained measures of the Flesch formulas. At the same time, however, it must be remembered that the entire set of reports has demonstrated a great degree of homogeneity and narrow range under the Flesch technique.

The reports of railroad companies have the greatest amount of variability on the measures employed. Their range in average sentence length is from 18 words (*standard*) to 53 words (*very difficult*). This observation seems to reflect the great difficulty attached to railroad reporting due in part to legal specifications concerning content. Apparently, however, some railroad companies are fulfilling their legal and public obligations in a more effective manner of readability than others.

Another striking inference may be obtained from Table 55.1. The arrangement by rank order in readability attainments under the formulas, parallels almost directly the degree of contact the companies have with the general public. Merchandising, Communications, Foods, Automobiles and Accessories corporations cater to larger sections of the general public. Corporations dealing in Machinery and Supplies, Metals and Chemicals have a more restricted, less diversified market for their products (8). Since this observation is based on the small and variable groups of the study, however, more data and analysis are actually required for final proof.

Nevertheless, such an arrangement is by no means unwarranted. It is generally accepted that the extent and character of the public interest should first be determined in the construction of the report. Then a corporate writer attempts to write to that audience. However, if this same trend had appeared within lower degrees of difficulty, it could be considered a more legitimate consequence of the nature of the enterprise and the groups interested in its operations. At the same time, it would then meet the readability requirements of these particular audiences.

Summary

1. Analysis of the readability of the twenty-six annual reports of corporations listed in the Billion Dollar Club of *Business Week*, June 11, 1949, by means of the new Flesch Readability Formulas, revealed that, on the whole, the general level of reading was *difficult*, and the human interest value *dull*.

2. These reports contain language which is beyond the language experience and fluent comprehension of approximately 75 per cent of the U. S. adult population.

3. The Flesch technique demonstrates promise as a method for indicating the difficult language elements in corporate reports.

4 It also demonstrates promise as a method for spotting 'impersonalness' in such writing

When, as in this study, the writing sample involves a problem of mass communication, the Flesch technique appears to be a reasonable instrument It gauges the likelihood that these annual reports will convey their messages to most of their prospective readers Wider application of the technique in the construction of the annual report is recommended Used in conjunction with the other types of practical hints in the literature, it can serve to strengthen the annual report as the most important single written communication between management and stockholders, employees, and the general public

REFERENCES

1 Dale E, *Preparation of Company Annual Reports* Research Report No 10 American Management Association, New York, 1946, 104 pp
2 Doris, Lillian, *Modern Corporate Reports* New York Prentice Hall, Inc, 1948
3 Flesch, R F, *Marks of Readable Style A Study in Adult Education* New York Bureau of Publications, Teachers College Columbia University, 1943 (Contributions to Education, No 897)
4 Flesch, R F, *The Art of Plain Talk* New York Harper and Brothers, 1946
5 Flesch, R F, A New Readability Yardstick, *Journal of Applied Psychology* 1948 Vol 32, 221-233
5a Flesch, R F, *The Art of Readable Writing* New York Harper and Brothers, 1949
6 Flesch, R F, Making the Narrative Readable Chapter 15 in *Modern Corporate Reports* by Lillian Doris, New York Prentice Hall Inc, 1948
7 Gibson, W B ed, *The Annual Report* A study of over 500 financial reports of leading American Business Institutions showing the present style trend and important physical characteristics Chillicothe, Ohio Mead Corporation Marketing Research Division, 1939
8 McLaren N L *Annual Reports to Stockholders—Their Preparation and Interpretation* New York Ronald Press, 1947

Readability and Human Interest of Management and Union Publications *

JAMES N FARR, DONALD G PATERSON, and C HAROLD STONE

The authors have benefited from the advice and assistance of Dr Dale Yoder, Director of the Industrial Relations Center, and of Dr Herbert G Heneman Jr, Assistant Director

The Industrial Relations Center Reference Room regularly receives and files a wide assortment of company house organs and labor union newspapers and journals A random sample of twenty five management publications and twenty five union publications was drawn for this study

[EDITOR S NOTE Last year Flesch s *The Art of Readable Writing* was published primarily, it seems, to give editors of "learned journals a bad conscience Academicians are now trying to spread the bad conscience to people who write for management and union publications This article from the Industrial Relations Center of the University of Minnesota demonstrates that learned journals are not the only publications that fail to meet Mr Flesch s high or low literary standards Whether Mr Flesch s efforts will result in reducing high thinking to the level of low living remains to be seen James N Farr was a member of the

* Reprinted from *Industrial and Labor Relations Review* Vol 4, No 1, October 1950

Industrial Relations Center staff at the time that the study was made and is at present an instructor of psychology at the University of Minnesota Donald G Paterson is Professor of Psychology and member of the Industrial Relations Research staff, University of Minnesota C Harold Stone is Assistant Professor of Psychology at the University of Minnesota and Research Associate at its Industrial Relations Center]

Are management publications and union publications written at a sufficiently simplified level to ensure ready understanding by rank and file employees and members? If the answer is No, then much of the effectiveness of these publications is nullified Unfortunately, evidence reported in this article strongly supports a No' answer

The evidence was secured by applying the Flesch formulas for measuring readability of written communications to 25 management house organs and to 25 union newspapers The Flesch formula measures readability by taking into account the average length of sentences used and the average length of the separate words used[1] Thus, long and involved sentences are penalized Short, simple sentences are rewarded In similar manner, excessive use of multi syllable words (jawbreakers) is penalized The use of short, single syllable words, on the other hand, is rewarded The result is a scale ranging from 0 (very difficult) to 100 (very easy)

Flesch provided a table for interpreting these scale scores in terms of the previous education (school grades completed) required for ready understanding Table 56 1 gives the essentials

Inspection of Table 56 1 suggests that written communications intended for the overwhelming proportion of rank-and file employees should secure a reading ease score above 70

The discrepancy between measured reading ease of union newspapers and management house organs[2] and the requirements

TABLE 56 1

Flesch s Interpretation of Reading Ease Scores

Reading ease score	*Description*	*Education required for understanding*
0–30	Very difficult	College
31–50	Difficult	High school or some college
51–60	Fairly difficult	Some high school
61–70	Standard	7th or 8th grade
71–80	Fairly easy	6th grade
81–90	Easy	5th grade
91–100	Very easy	4th grade

set forth in Table 56 1 is startling The range of reading ease scores for union newspapers and for management house organs is shown in Chart 1 Each x' represents one publication A glance at Chart 1 shows that not one of the fifty publications receives a reading ease score above 70 Furthermore, it is evident that, on the whole, the union newspapers are written at a more difficult level of readability than are the house organs

TABLE 56 2

Mean Reading Ease Score of Management and of Union Publications *

	Number studied	*Mean reading ease*	*S D*	*Interpretation*
Management publications	25	52	8 8	Fairly difficult
Union publications	25	40	10 0	Difficult

* Range for management, 39 to 69 for union, 21 to 60

Table 56 2 gives the average reading ease score for the management and for the union publications, together with an index of variability (called ' standard deviation by statisticians) This table, read in conjunction with Table 56 1, indicates that, on the average, these publications are pitched at a level suitable for employees with a high school or a college education[3]

[1] R Flesch, *The Art of Readable Writing* New York Harper and Brothers 1949, p 237

[2] The same procedure was used in studying each of the 50 publications Samples of 100 words each were taken randomly from each publication until a minimum of 10 per cent of the written material had been covered

[3] The Flesch reading ease score for the present paper is 40 or difficult Average

The reason for the discrepancy is not hard to understand Editors of union newspapers and of company house organs are, for the most part, college educated or self educated to the college level It is, therefore, second nature for them to write at a high brow level without realization of that fact An example will illustrate the point A former union business agent who became a factory personnel manager asked the Minnesota Industrial Relations Center to Flesch" an employee handbook which he had written He was astonished to get the report that his handbook was written at the very difficult level He protested that this could not be because he had left school at the end of the eighth grade But he was reminded of the fact that he was a great reader and a serious student of industrial relations literature Furthermore, he had enrolled in a graduate school extension course in industrial relations and had made an excellent scholastic record in competition with graduate students It then dawned on him that he was really different from his former fellow workers who likewise had had only an eighth grade education but had continued as rank and file employees at the semiskilled level of work

Human Interest Level

Flesch points out that another dimension of written communications must also be measured He refers to the fact that printed material should be written in story form, patterned after the way we actually talk to one another If it is so written, it will have greater attention holding power This is the 'human interest' element, and it can be measured by the proportion of sentences that contain "personal references, such as names of people, and direct quotations It can also be measured by the percentage of personal words such as pronouns included in a 100 word sample By taking these two factors into account, the result is a scale ranging from 0 (dull) to 100 (dramatic)

Table 56 3 gives the Flesch interpretations for the human interest scale It is apparent that writers desiring to ensure maximum attention should strive to have their copy reach a score of 41 or more

TABLE 56 3

Flesch s Interpretation of Human Interest Scores

Human interest score	*Description of style*	*Typical magazine*
0–10	Dull	Scientific journal
10–20	Mildly interesting	Trade journal
20–40	Interesting	Digests
40–60	Highly interesting	*New Yorker*
61–100	Dramatic	Fiction

Again there is a startling discrepancy between measured human interest of union newspapers and management house organs [4] and the desired human interest levels as set forth in Table 56 3 The range of human interest scores for union newspapers and for management house organs is shown in Chart 2 Again, each x represents one publication Not one of the publications reaches the 'dramatic level and only two reach the 'highly interesting level None of the union newspapers reach these two desirable levels Furthermore, the majority of both house organs and union newspapers are only 'mildly interesting or dull

Table 56 4 gives the average human interest score for management and union publications, together with an index of variability (standard deviation) This table, read in conjunction with Table 56 3, indicates that, *on the average* these publications fall far short of the highly interesting' or dramatic' levels of human interest that would be desirable

Additional detailed evidence, not re

sentence length, however, is only 18 or 'standard but the average number of syllables per 100 words is 177 or difficult In other words, the present paper is relatively easy to read so far as brevity of sentences is concerned, but it moves into the difficult level because of the number of three and four syllable words contained in it

[4] The Flesch formula for measuring human interest was applied to the same 100 word samples taken from each of the fifty publications that were used in measuring reading ease

TABLE 56 4

Mean Human Interest Score of Management and of Union Publications *

	Number Studied	*Mean Human Interest*	*S D*	*Interpretation*
Management publications	25	22	10 4	Mildly interesting
Union publications	25	15	6 4	Dull

* Range for management 6 to 45, for union, 6 to 33

ported here, shows that, for those sections of both management and union publications dealing with general reporting, news, and editorials, the human interest scores hover around the lower levels of human interest Personals' in both types of publications reach the 'interesting level If it were not for the presence of personals in both types of publications, they would sink to a still lower level of human interest

It is evident that editors of union newspapers and of company house organs can and should greatly improve their publications by following Flesch s rules for simplifying written language and for increasing the interest value of what is written

Nothing in the present report bears on the virility or dynamic value of the ideas' contained in the two types of publications Only a content analysis would disclose this It is the impression of the writers that such a 'content analysis would demonstrate an important difference The company house organs would probably be characterized as pale, anemic, and

CHART 1

Distribution of Reading Ease Scores of Union and Management Publications *

Union Newspapers

0–30 Very diff	31–50 Diff	51–60 Fairly diff	61–70 Standard	71–80 Fairly easy	81–90 Easy	91–100 Very easy
	x					
	x					
	x x					
	x x					
	x x					
	x x					
x	x x					
x	x x	x				
x	x x	x				
x	x x	x				

Management House Organs

0–30 Very diff	31–50 Diff	51–60 Fairly diff	61–70 Standard	71–80 Fairly easy	81–90 Easy	91–100 Very ea y
	x					
	x					
	x	x				
	x	x				
	x	x				
	x	x	x			
	x	x	x			
	x	x	x			
	x x	x	x			
	x x	x	x			

* Each x represents the reading ease score of one publication

CHART 2

Distribution of Human Interest Scores of Union and Management Publications *

Union Newspapers

0–9 Dull	10–19 Mildly interesting	20–39 Interesting	40–59 Highly interesting	60–100 Dramatic
	x			
	x			
	x			
	x	x		
	x	x		
x	x	x		
x	x	x		
x	x x	x		
x	x x	x		
x	x x	x		

Management House Organs

0–9 Dull	10–19 Mildly interesting	20–39 Interesting	40–59 Highly interesting	60–100 Dramatic
		x		
	x	x		
	x	x		
	x	x		
	x	x		
	x	x		
x	x	x		
x	x	x		
x	x	x	x	
x	x	x	x	

* Each x represents the human interest score of one publication

generally lacking in the kind of information that Heron has forcefully suggested employees demand[5] Union publications, on the other hand, would probably be found to be red blooded and full of demands for better wages, cost of living data, and iteration of grievance promoting situations Regardless of these probable differences the present study indicates that management editors and union editors alike need to work strenuously toward language simplification and enhanced human interest values If they do this their publications will be readily understood by rank and file employees with a limited educational background

[5] A R Heron, *Sharing Information with Employees* Stanford, California Stanford University Press, 1940, p 204

Chapter XIV

FORCED CHOICE AND CRITICAL REQUIREMENTS

Advancement in knowledge is obtained not only from continued use of known techniques and concepts but also from variations, resulting in new and different techniques When such techniques evolve, there are always those who pioneer

and offer enthusiastic support Others look upon the technique with skeptical criticism There is need for both views Without the energetic and enthusiastic endeavors of the former, science would remain either static or move impercep tibly Without the skepticism of the latter, the techniques so challenged would not be sharpened and improved Ultimately, the proposed technique, if valid, stands the tests of time and becomes generally acceptable, or its shortcomings be come sufficiently obvious and it falls into disrepute

Reviewing the discoveries and techniques in such fields as medicine, child psychology, and psychological testing presents adequate examples One need think only of the Sister Kenny treatment of polio, child feeding schedules, and the variety of projective tests beginning with the Rorschach to see clearly that although newer proposals result in controversy, when the dust clears both the protagonists and antagonists have contributed to the advancement of the concept or technique

A field as relatively young as experimental industrial psychology can be expected to experience tremors as a result of new proposals The two techniques included in this chapter are Forced Choice and Critical Requirements A sep arate chapter on job or employee evaluation has not been included in this book, but characteristic of such evaluation systems have been various types of ratings The two techniques presented can be considered as alternative techniques at tempting to avoid some of the defects of the older rating scales and methods

The "Forced Choice" technique requires a rater to choose one of two desirable adjectives or phrases which describe the person to be rated The rater must also choose one of two undesirable qualities Although both items seem either equally desirable or undesirable, only one of each pair discriminates between competency and incompetency Sisson's article explains how the forced choice rating evolved It gives examples of the tetrads used, in addition to presenting a technical dis cussion of the details in construction of the items

The Industrial Relations Memo (119) is a report of the use of this technique with supervisors in the Esso Standard Oil Company In addition to presenting a step-by-step account of the research, the last section is concerned with the merits and limitation of the technique

Travers is critical of forced choice, and has published a review in which he raises very important issues concerning the rationale of the technique as well as its validity Baier responded to Travers' critical article With both authors placed on the record, it is possible for the reader to evaluate the arguments and decide for himself Characteristic of many psychological periodicals is to allow contro versy in print This is a very favorable and wholesome thing calculated to improve standards and increase knowledge The one certainty is that the technique will be both more widely used and criticized Eventually the answer will be known more clearly than it is at this time

Another technique has been proposed by a group of workers under the leader ship of John C Flanagan It is known as "Critical Requirements," and may be considered a new approach to employee evaluation It has also been used to identify sources of difficulties experienced by operators in using equipment and in other ways illustrating the versatility of the technique

There are differences between the critical requirements technique and the method of forced-choice, but there are also similarities Both evolved during World War II, one in the Army and the other in the Air Force Both are con-

cerned with evaluating personnel, one primarily from the *man* aspect and the other emphasizing the *job* aspect Both obtain data by emphasizing information that differentiates competency from incompetency, one in terms of ratings of be havior and the other in terms of specific job incidents Flanagan's article, "Critical Requirements—A New Approach to Employee Evaluation' describes the tech nique as it applies in the evaluation of Air Force officers The Gordon article applies the technique as a method of evaluating flying skill and indicates the possibility of application to a wide variety of problems No critical and scholarly review of this technique has, so far, been published, but it will be extremely unusual if such articles do not appear soon

Forced Choice—The New Army Rating *

E DONALD SISSON

The research reported in this study represents the combined efforts of the entire profes sional staff of the Personnel Research Section, AGO in 1945 The opinions expressed, how ever, are those of the author and do not necessarily represent those of the Department of the Army

SUMMARY

The origin of the use of efficiency re ports for Officers of the U S Army is lost in history, as is the story of the evolution of the formal procedures of reporting Sometime after the first World War, how ever, a standard form was adopted and a procedure regularized for accomplishing this report Thereafter, twice each year—on June 30th and on December 31st—every officer in the Army has been rated by his immediate superior, and this rating submitted to the War Department Though early recognized as not completely satisfactory, the original rating form remained in force (with sporadic minor amend ments) until it was superseded in July of 1947

The new form is the product of many months of concentrated research It is rad ically different in many respects from the old form, and from other rating devices currently in use in industry Its most novel feature is the use of what has been called the 'forced choice" rating method Rather than indicating how much or how little of each characteristic an officer possesses, the rater is required to choose, from several sets of four adjectives or phrases, which best characterizes the officer, and which is least descriptive In other words, it calls for objective reporting and minimizes sub jective judgment And because of the way in which the tetrads—sets of four rating elements—are constructed, it reduces the rater s ability to produce any desired out come by the choice of obviously good or obviously bad traits It thus diminishes the effects of favoritism and personal bias

The technique, and the form embody ing it, has been tried out on fifty thousand officers—in both experimental and official trials—and the results obtained with it have been compared with independent criteria of efficiency arrived at through group ratings The new method is superior to all other methods examined It produces a better distribution of ratings relatively free from the usual pile up at the top of the scale It is less subject to in fluence by the rank of the officer being rated It is quickly and objectively scored by machine And above all, it produces

* Reprinted from *Personnel Psychology* Vol 1, No 3 Autumn 1948

ratings which are more valid indices of real worth

The particular form developed for rating Army officers would probably be of little value for other groups—largely because of the specificity of the rating elements it contains The technique, however, has already proved of value in other situations and there is every reason to believe that it is even more generally applicable

The Old Rating System

It can generally be assumed that the main value of efficiency ratings—usually their sole purpose—is in providing a sound basis for personnel actions Yet when the clouds of war rolled up in 1940 and it became evident that the Army needed to promote a rather large number of top ranking officers immediately to serve as generals of the rapidly mobilizing forces, it was suddenly discovered that the years of regular efficiency reporting had provided no basis for the important decisions that had to be made To quote one of the men responsible for making the selections at that time Efficiency reports, instead of showing the 150 best, showed only that of 4000 ground officers of suitable general officer age, 2000 were superior and best As such a showing was perfectly worthless for the purpose, the selecting authorities reluctantly fell back on personal knowledge, which is exactly what the Army thought it was getting away from when, 20 years ago, it inaugurated the existing system"

The existing system was not as bad as this recital might make it appear As such systems go, in fact, it was fairly typical and quite respectable, it would even compare rather favorably with the run of the mill of systems currently in use in business and industry It contained some ten numerical scales covering such general traits as 'force, "leadership,' "attention to duty, 'ability to obtain results," and so forth, and each scale, as well as the net numerical score, was divided into areas which were assigned the five adjectival ratings of superior, excellent, very satisfactory, satisfactory and unsatisfactory It was generally filled out with great care, and undoubtedly with great seriousness—nothing is more important to the Army officer than his efficiency index Moreover, its validity as determined through extensive tests was shown to be at least fair, particularly with respect to the identification of a very small number of outstandingly poor officers

If a superior officer really and honestly wanted to point up the deficiencies in a truly poor subordinate, the form was adequate to the purpose But therein lay its greatest weakness, the rater could control the outcome at will And because of traditions, the pressures of circumstances and for a host of other reasons—personal or general—he usually made it come out high He said only the best of his men or else damned with faint praise' by saying the next best about those whose performance was low Or if his conscience pricked, he said nothing, and left the trait unrated with a cryptic "unknown" Nothing but good' was the general rule, with the consequence that the whole scale was distorted, what was supposed to be outstanding became typical, and to be labelled satisfactory' was to be called intolerably inefficient To correct these deficiencies in the system, and to provide a more valid procedure for rating Army officers, work was begun in 1945 on the development of a new efficiency report

The New Officer Efficiency Report

At the outset the research leaned pretty heavily on the finding of a recently completed program for screening war time officers to be offered commissions in the Regular Army One of the instruments developed in that program—and shown to possess a high degree of validity—was a rating form which incorporated, among other elements, a forced choice section In addition, a by product of the earlier research program—a method for constructing an acceptable criterion—was of equal importance to the present problem This latter will be discussed first

The criterion The crucial importance of the criterion in research of this kind cannot be over emphasized To determine

the validity of any rating system, it is obviously necessary to compare ratings produced by it with some independent measure of each man's true merit In this case, the criterion problem was attacked by identifying groups of officers who were clearly outstanding in efficiency or competence and other groups clearly less competent This identification, of course, could not be based on existing efficiency reports since to do so would beg the whole question Nor could it be based solely on other such opinions of superiors It was decided, therefore, to use the consensus of fellow officers in identifying the Army's best and poorest The procedure followed was somewhat as follows Officers belonging to the same unit, and in a position to know each other's work and qualifications, were assembled in groups of about 20 to 40 Each was furnished a form on which all names appeared in alphabetical order, regardless of rank Without signing the form, or identifying himself in any way, each officer was asked to select the best—most competent—of the group, then the least competent, and to continue selecting most and least competent until all but about five names on the list had been selected as among the most or least competent By tallying these nominations, it was possible to earmark the two or three in each group who were clearly best, the two or three clearly poorest, and finally, from among the names not rated either high or low, some truly average officers By repeating the process in literally hundreds of such units, comprising almost 50,000 officers, rather sizable groups of high, middle, and low officers were identified

Members of these three widely divergent criterion groups were rated in the normal manner on various types of rating forms Needless to say, no rating officer was apprised of the criterion status of the ratee Results obtained with these independent ratings were correlated with criterion group membership In all comparisons, one particular form stood out as most valid This was the form containing the forced choice elements mentioned above

How forced choice items are made Forced choice rating elements are sets of four phrases or adjectives pertaining to job proficiency or personal qualifications The rater indicates which of the four is most characteristic of the ratee and which is least characteristic, and repeats this selection for each of the sets included A sample set is the following

A Commands respect by his actions
B Coolheaded
C Indifferent
D Overbearing

It is at once obvious that two of these are relatively favorable terms and the other two relatively unfavorable One of the two favorable terms, checked as most characteristic, gives plus credit, selecting the other gives no credit In the same way, picking one of the two unfavorable items as least characteristic adds credit whereas the other adds nothing

The construction of these tetrads and the determination of the scoring key are the crucial problems in the development of a rating scale of this type Rundquist (2) outlined 6 steps in the process

1 Collection of brief essay descriptions of successful and unsuccessful officers
2 Preparation of a complete list of descriptive phrases or adjectives culled from these essays, and the administration of this list to a representative group of officers
3 Determination of two indices for each descriptive phrase or adjective—a preference index and a discrimination index
4 Selecting pairs of phrases or adjectives such that they appear of equal value to the rater (preference index) but differ in their significance for success as an officer (discrimination index)
5 Assembling of pairs so selected into tetrads
6 Item selection against an external criterion and cross validation of the selected items

The New Officer Efficiency Report, as it was approved for official use, consists of 12 of these forced choice tetrads relating to job proficiency, followed by two 10 point graphic scales concerning the ratee's primary and secondary duties (Fig 57 1) Then there are 12 more tetrads pertaining

to personal qualifications, followed this time by six 10 point scales concerning such general characteristics as cooperation—spelled out as 'The degree to which he is able and willing to work with other officers and enlisted men —or initiative, the "degree to which he is able to act on his own responsibility in the absence of orders ' (Fig 57 2) These sections, which constitute the scorable part of the form, are printed on an IBM answer sheet Preceding this, and attached to it but perforated to permit easy detachment, is a sheet calling for identifying information, a verbal description, recommendations and other information of an administrative nature (Fig 57 3)

The "For Keeps" trial As already indicated, various preliminary forms of this report were tried out experimentally, validated against the criterion described above, and compared with other types of reports In all of these experiments involving the experimental rating of almost 50,000

EFFICIENCY REPORT

WD AGO Form 67 1 Part 2 See AR 600–185 for details.

Unit Adjutant or Personnel Officer will complete Sections I and III.
Rating Officer will compl te Sect ons II, IV V VI, VII VIII, and IX.
Indorsing Officer will complete Sections II, V VII and IX.

Section III OFFICER REPORTED UPON
Enter same information as for Section I

LAST NAME FIRST NAME INITIAL	SERIAL NUMBER	GRADE	ARM OR SERVICE	COMPONENT	PERIOD OF REPORT FROM	TO	DO NOT WRITE IN THIS SPACE

THEATER OR CONTINENTAL COMMAND	UNIT ORGANIZATION, AND STATION	PRIMARY MOS	DUTY ASSIGNMENT (MOS CODE)	DAYS OF DUTY	LEAVE	OTHER NON-DUTY	JP

DATE OF REPORT	FOR REPORTS RENDERED BECAUSE OF PERMANENT CHANGE OF STATION, SUPPLY ADDRESS OF UNIT AND INSTALLATION WHERE OFFICER WILL REPORT	PQ

OA

READ INSTRUCTION SHEET CAREFULLY BEFORE MARKING THIS SECTION

Section IV JOB PROFICIENCY

1. A. Becomes dogmatic about his authority B. Careless & slipshod in attention to duty C. No one ever doubts his ability D. Well-grounded in all phases of Army life. — MOST / LEAST
2. A. Follows closely directions of higher echelons. B. Inclined to gold-brick. C. Criticizes unnecessarily D. Willing to accept responsibility — MOST / LEAST
3. A. A go-getter who always does a good job. B. Cool under all circumstances. C. Doesn't listen to suggestions. D. Drives instead of leads. — MOST / LEAST
4. A. Always criticizes, never praises. B. Carries out orders by "passing the buck." C. Knows his job and performs it well. D. Plays no favorites. — MOST / LEAST
5. A. Constantly striving for new knowledge and ideas. B. Businesslike. C. Apparently not physically fit. D. Fails to use good judgment — MOST / LEAST
6. A. Cannot assume responsibility B. Knows how and when to delegate authority C. Offers suggestions. D. Too easily changes his ideas. — MOST / LEAST
7. A. Fails to work for the best interest of all. B. Has high degree of initiative. C. Never makes excuses for his mistakes. D. Slow in accomplishing his work. — MOST / LEAST
8. A. Criticizes policies of superiors. B. Others can't work with him. C. If he is wrong, will admit it. D. The men know they can rely on his judgment — MOST / LEAST
9. A. Doesn't try to "pull rank. B. Knows men, their capabilities & limitations. C. Low efficiency D. Uses steady monotone in his speech. — MOST / LEAST
10. A. Fails to support fellow officers. B. Oversteps his authority C. Gives clear and concise directions. D. Very exacting in all details. — MOST / LEAST
11. A. Blames others for his mistakes. B. Always demands strict discipline. C. Excellent at constructive criticism. D. Hesitant about rendering decisions. — MOST / LEAST
12. A. Can take over in an emergency B. Fair and just in his dealings. C. Lacks interest in his job. D. Questions orders from superiors. — MOST / LEAST

READ INSTRUCTION SHEET CAREFULLY BEFORE MARKING THIS SECTION

Section V JOB PROFICIENCY

DO NOT WRITE IN THIS SPACE

1. Management and operation of military matters not included in tactics and strategy
2. The directions of the over-all operation of military unit.
3. Presenting learning materials in classroom situation in military or civilian component
4. Exercise of specialized knowledge, requiring lengthy technological training.
5. Assisting commanders of battalions or larger units in devising methods of meeting the requirements of military situations.
6. Duties involving aeronautical skills performed by rated officers.
7. Training at service schools, Air University Army Industrial College, etc.

	FOR RATING OFFICER		FOR INDORSING OFFICER
PRIMARY	1 4 5 6 8 0	PRIMARY	1 2 3 4 5 6 7 8 9 10
SECONDARY	8	SECONDARY	1 2 5 6 7 8 9 0

DO NOT WRITE IN THIS SPACE

FIGURE 57 1 *Job proficiency section of new officer efficiency report*

officers the new report proved more valid and more acceptable in several other respects But since these were experimental trials, in which the pressures and circumstances surrounding official reports were not called into play, it was recommended that a real test be made in an official reporting period Consequently both the old form (known as Form 67) and a later version of the new report (to be labelled Form 67 1) were used on all officers throughout the country for the regular semi annual efficiency report in June 1946 At the same time, but with precautions to prevent cross contamination, criterion information was collected on fifteen thousands of these same officers

Analysis of the data obtained in these studies led to the conclusion that the new form (67 1) was clearly superior to the

F 1429 RE 4

Section VI PERSONAL QUALIFICATIONS

Use ELECTROGRAPHIC PENCIL, foll w g s me d ec ons s fo S ct IV M k ONE m k EACH c lum f r ea h set f tems

Set	Items
1	A. P pl w k f & w h h m be se f h pe sonal y; B. N k s; C. Th ks ly f h ms lf; D. W g ld l
2	A. A t thle s; B. F m but t erbea ng; C. Eg t t L; D. Rub peopl th w g w y
3	A. C mpl me man on h good work; B. Lose h h ad get e ted; C. Ha adm at on of officers & men l ke; D. Poor dres & appearance
4	A. La k abil ty t fid f me & ll rs; B. E yg g; C. Type f man ev ry l kes f fri nd; D. Ha a qu t d g f ed be g
5	A. Hot t mper d; B. F l t d mo t le or g nal y; C. Reserved; D. Impresses p pl favor bly
6	A. Boastf l; B. I p es pr de in the organization; C. Lacks tact; D. Thoughtful of others
7	A. Pl ty f ml y sn p bea g & eat; B. Normally h rf l; C. Ca t t ke cr t sm; D. D t get l g w th peopl
8	A. Modest & served; B. Doe h d or f h hould; C. A t social; D. Respected by all f ll w ll er
9	A. Aq el, unassuming ff; B. F llows ather th lead; C. Has an attit d of superior ty; D. Tactful
10	A. Obt pect & obed en w h ut g sentm; B. L k ggre en ss; C. H ellent ommand l l g; D. La k g good ondu t & mor l habi
11	A. C l aded; B. C mma d r spect by h ac; C. Ov bear g; D. I d ll nt
12	A. Imma ure; B. Modest but not ret ng; C. N rvous; D. Thor gh l ooperat h work

Each set is marked M (most) and L (least) against A, B, C, D.

Section VII PERSONAL QUALIFICATIONS

Use ELECTROGRAPHIC PENCIL, follow ng same d ectio s as for Section V MARK ALL SIX QUALIFICATIONS

	FOR RATING OFFICER	FOR INDORSING OFFICER
The degree to which he is able to meet tuations without bias and without emotional upset	1 2 3 5 6 7 8 1	2 3 5 10
The degree t wh h he bl nd w ll g t work w th other officer and enl ted men.	1 3 5 8 9 10	1 2 3 5 6 7
The deg ee t which he bl t act on h own respon bility in absenc of ord	1 2 3 4 5 6 7 9 0	2 3 5 6 0
The degree t wh ch he is abl t d scr minat & ev luat fact t arr at log cal con lusions.	1 3 5 6 7 9 0	3 5 6
The deg ee t which h appearanc nd beha ior cause peopl to react favor bly	1 2 3 4 5 6 7 9	1 2 3 5 8 1
The degree t wh h he bl t carry out order with consistency & fi mness t achieve object ves,	1 2 4 5 6 7 8 0	1 2 3

Section VIII OVER ALL RELATIVE RANK

FOR RATER ONLY

The number of off ers th s grad ed by me at this t me is______	If these ff ers wer rr ged d c ider ng over ll f tu sef l ness t the Army f om h gh st (N 1) t poorest th ll er w ld be No.______ f the t tal g p ed

Section IX AUTHENTICATION

Use typew t (ept f g t re) or nk.

I cert fy that I have read the curre t AR 600-185 and that all rat gs ar made in ccordance with i struct o s conta ned th a d that to the best of my knowledg and bel f all t es c ta ned h reo a e tru a d mp rt al

SIGNATURE OF RATING OFFICER	SIGNATURE OF INDORSING OFFICER
NAME, GRADE, AND ORGANIZATION OR UNIT	NAME GRADE, AND ORGANIZATION OR UNIT
OFFICIAL STATUS OF RATED OFFICER WITH RESPECT TO RATING OFFICER	OFFICIAL STATUS OF RATED-OFFICER WITH RESPECT TO INDORSING OFFICER

FIGURE 57 2 *Personal Qualifications Section of New Officer Efficiency Report*

EFFICIENCY REPORT

See AR 600-185 for details.

Unit Adjutant or Personnel Officer will complete Sections I and III.
Rat ng Officer will complete Sections II IV V VI, VII, VIII, and IX.
Indors ng Officer will complete Sections II, V VII, and IX.

Section I OFFICER REPORTED UPON

Use typewriter or print in ink. Use carbo pape to f ll out Sectio III at same t me. See AR 600-185

LAST NAME FIRST NAME INITIAL	SERIAL NUMBER	GRADE	ARM OR SERVICE	COMPONENT	PERIOD OF REPORT FROM	TO	DO NOT WRITE IN THIS SPACE

THEATER OR CONTINENTAL COMMAND	UNIT ORGANIZATION, AND STATION	PRIMARY MOS	DUTY ASSIGNMENT (MOS CODE)	DAYS OF DUTY	LEAVE	OTHER NON-DUTY	JP

DATE OF REPORT	FOR REPORTS RENDERED BECAUSE OF PERMANENT CHANGE OF STATION, SUPPLY ADDRESS OF UNIT AND INSTALLATION WHERE OFFICER WILL REPORT	PQ

NAME, GRADE, AND ORGANIZATION OR UNIT OF RATING OFFICER	NAME, GRADE, AND ORGANIZATION OR UNIT OF INDORSING OFFICER	QA

Section II DATA AND SUGGESTIONS FOR USE IN ASSIGNMENT

NOTE Informatio thi pag will be f rwarded t the Career B nch of the Personnel nd Admi istration Division by TAG after ratings have been determined. Proper future assig ment nd ut l zati of the officer will depend upon the care with which information in this section is formulated and reported. Use typewriter or print in ink.

A. DUTIES ACTUALLY PERFORMED ON PRESENT JOB To be suppl ed by Rater Be specific. G ve his duty assignment and all additional duties w th enough specific deta l to show scope of job n each area.

B. DESCRIPTION OF OFFICER RATED AND COMMENTS. These parag aphs should cover physical mental, moral qual t es of rated off er specialties of alue to the Army and any special d fects or we knesses affecting his ab l ty to do certa n assignments.

COMMENTS OF RATING OFFICER	COMMENTS OF INDORSING OFFICER

C. ESTIMATED DESIRABILITY IN VARIOUS CAPACITIES. Assume you are a comm der of a major un t n war Ind cate to what extent you would want the ated off cer to serve under you in the ext h gher grade in each type of duty descr bed below Place a X the proper box using the sh ded NA area if the duty is not appl cable If l ne h is used spec fy the nature of the pecialty

	Rater NA	Rater 1 NOT WANT HIM	Rater 2 TAKE A CHANCE ON HIM	Rater 3 HAPPY TO HAVE HIM	Rater 4 PREFER HIM TO MOST	Rater 5 FIGHT TO GET HIM	Indorser NA	Indorser 1 NOT WANT HIM	Indorser 2 TAKE A CHANCE ON HIM	Indorser 3 HAPPY TO HAVE HIM	Indorser 4 PREFER HIM TO MOST	Indorser 5 FIGHT TO GET HIM
a Represe t your viewpo t and make dec s ns y ur ame t a h gher h dq arters.												
b Command a unit immed tely subo d te to yo on omb t m on												
c. Be espons ble an emergen y call ng fo t t ve cool o s forceful leadersh p												
d W rk o an ass gnme t req ring g eat ttent on to det l and out ne												
e Pl ll spects of a m l t ry tuat n u ng judgme t in t at o a d coolness.												
f Ca ry o t an ass gnme t h a civ l an compone t such as ROTC NG or ORC.												
g Represent you where tact and abil ty t get al ng with people are needed.												
h Wo k o an ass g ment as pecialist or tech cia (Spec fy)												
i. Carry out the dut es of the type of work to wh ch he is now as gned.												

D IMMEDIATE RECOMMENDATIONS FOR CAR ER DEVELOPMENT Be specif c.

RATER'S RECOMMENDATION FOR ASSIGNMENT (MOS CODE)	INDORSER'S RECOMMENDATION FOR ASSIGNMENT (MOS CODE)
RATER'S RECOMMENDATION FOR FURTHER TRAINING	INDORSER'S RECOMMENDATION FOR FURTHER TRAINING

E. ENTRIES ARE BASED ON → (RATER WILL CHECK)	INTIMATE DAILY CONTACT	FREQUENT OBSERVATION OF THE RESULTS OF HIS WORK	INFREQUENT OBSERVATION OF THE RESULTS OF HIS WORK	ACADEMIC RECORDS	OFFICIAL REPORTS

WD AGO FORM 1 JUL 47 67-1 PART 1 — FORM I.T.S. 1100 F 1429 REV.

FIGURE 57 3 *Front sheet of new rating form incorporating personal information*

old form (67) This conclusion was based on several particulars

1 The new form produced ratings which were definitely less influenced—less biased—by the rank of the rated officer As a matter of fact, all ratings, and the criterion itself, were influenced by grade to some extent This is indicated in Figure 57 4, where the various scales are made comparable by converting each to standard score terms This figure shows the average score for the officers in each grade group from 2nd Lt through Colonel on the criterion[1] and on the various ratings Two

[1] The criterion, in this case was quantified by assigning values of 3 2 or 1 for each nomination of most competent, second most competent or any other high position, respectively, and values of −3, −2, or −1 for "nominations of least competent, next least competent or any other low rating by dividing by the number of nominators,"

ratings are shown for the new form One (labelled Section II, 67 1) is the score on the forced choice elements—not separated in this version into the two areas of job proficiency and personal qualifications The other (labelled Section III, 67 1) is the 'over all' rating—in this case a single 20 point scale on relative standing of the ratee in comparison with other officers *of his grade* The fact of influence by grade can be attributed partly to real differences (colonels in general are doubtless some what more efficient than second lieutenants in general because of the Army s promo tion policy) and partly to bias based on the prestige of rank In any event, the this advantage is more marked at the lower end of the scale, which means that the new form is particularly more effective in dis criminating among officers rated low in competence

3 Scores derived from both forms showed an unmistakable tendency to be higher when the ratings were rendered officially rather than in an experimental trial, but this tendency was much less marked for the scores on the new form than for those made on the old form 67

4 When scores on the two forms were compared with the independent criterion ratings of the same officers the new form was generally shown to be more valid

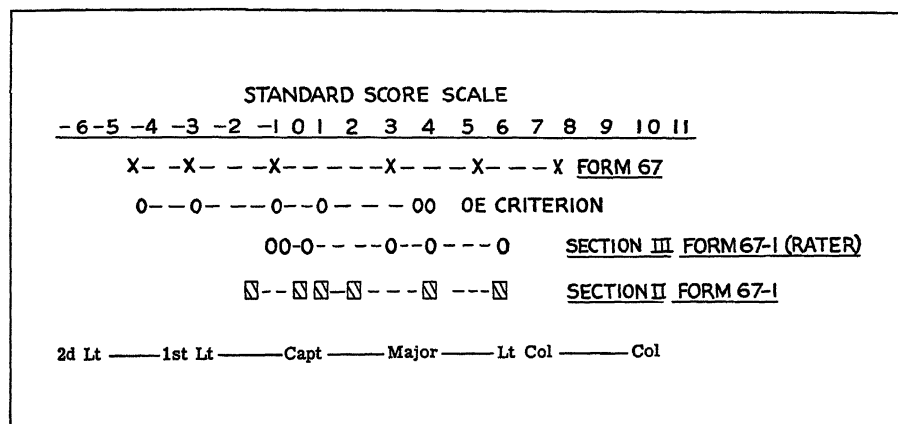



FIGURE 57 4 *Various scales made comparable by converting each to standard score terms*

older rating form (67) showed much more effect from rank than would be expected (more than the criterion index) while the two parts of the new form showed much less

2 Scores on the new form are distrib uted in a way which permits better dis crimination among officers rated at the two extremes of the scale In testing ter minology, the new form would be said to possess more floor and more 'ceiling" As indicated in Figure 57 5, which shows the actual distributions of scores on the two forms with range of scores equated,

After further revision along the lines al ready described, the form was again sub mitted to experimentation The results corroborated these earlier findings in every important respect In fact, its tested valid ity was demonstrated to be even higher than before After nearly two years of re search, it was felt that the form was defi nitely superior to any other yet devised and tested in fulfilling the requirements of an adequate rating system for Army pur poses—requirements outlined by General Witsell (3), The Adjutant General, in the following terms

It (an adequate rating system) should be capable of distinguishing between the

multiplying by 10 to clear decimals and adding a constant of 30 to avoid negative values

best and the next best in the Officer Corps instead of lumping them all together in the same category It should likewise indicate which are least efficient and which next least instead of merely labelling a micro scopic few at the bottom of the scale as unsatisfactory And finally, it should at long last and without fear or favor, admit that the average officer is truly average!"

Technical Discussion

Construction of the Forced Choice Tetrads As already noted, the scaling and selection of the rating elements to compose the forced choice tetrads is the nub of the problem The basic assumptions underlying the method can be stated as follows

1 Any real differences which exist between officers in competence or efficiency can be described in terms of objective, observable items of behavior

2 These behavior items' differ in the extent to which people in general tend to use them in describing other people, i e, in general favorableness,[2] and this tendency can be determined statistically

3 These items also differ in the extent to which they characterize officers at one extreme of the true scale of competence as opposed to officers at the other extreme The index of this difference, the "discriminative value, can also be determined statistically

4 Pairs of items can be selected such that they are equal in preference value but different in discriminative value A rater forced to say which item is most (or least) characteristic of a ratee is thus unable to select solely on the basis of prejudice for or against him (since the preference values are equal) The rater is compelled to consider both alternatives and—theoretically at least—to do a more objective job of reporting

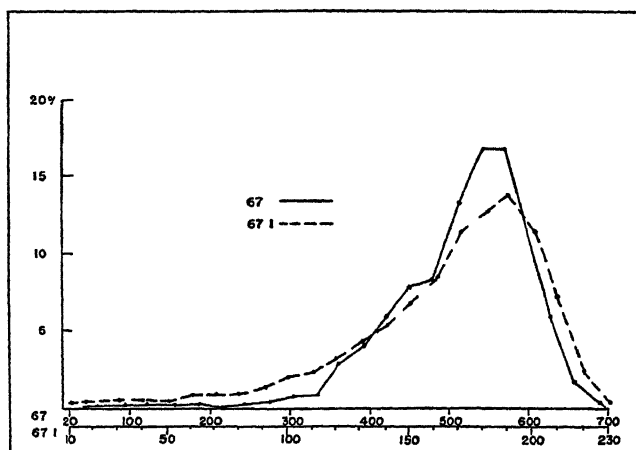



FIGURE 57 5 *Distribution of ratings on conventional rating scale Form 67 and distribution of ratings on the forced choice rating scale Form 67–1*

The first step in the process of constructing the tetrads of items, as stated above, is obtaining brief descriptive essays of good and poor officers These essays serve as the source of the behavior items pertinent to the job—in this case, the job of being an Army officer This step is essential, not only to focus agreement on the nature of the traits involved, but also to insure that the behavior items are worded in the language familiar to those who will later be using the scale

In the second step, a large pool of behavior items culled from these essays is

[2] Though not a necessity of the logic involved, those items which tend to be used most often, i e, are generally 'preferred by raters in describing others, are invariably more favorable items—nice things to say

prepared in list form and submitted to another group of officers—a group numbering in the neighborhood of 300 is generally used Each man in the group is asked to select from among his acquaintances some one officer whom he knows well enough to rate with confidence, and to indicate for that officer the extent to which each of the items in the list applies to him The following key is used for this purpose

1—to an *exceedingly high* or to the *highest possible* degree
2—to an *unusual* or *outstanding* degree
3—to a *typical* degree
4—to a *limited* degree
5—to a *slight* degree, or *not at all*

After completing the entire list in this fashion, each man is asked to evaluate the officer he is rating on a scale showing his position with respect to over all competence in a representative group of 20 officers of the same grade

All lists are collected, arranged in order of the rating of over all competence, and separated into upper (U), middle (M), and lower (L) thirds An analysis is then performed on each of the items, and a determination made for the three groups separately, of the frequency with which each of the five alternatives was chosen for that item Two values are then computed for each item

1 *Preference value* Assume that there are exactly 300 officers in the group checking the lists, and consequently 300 officers rated, divided into the three groups of 100 U, 100 M and 100 L For each item, the frequencies of each alternative, are summed across the three groups (U, M, and L) multiplied by the alternative weight (one less than the number preceding that alternative in the key presented above) and these five weighted alternative frequencies in turn added to yield a weighted total sum for the item This weighted total sum (which has limits of 0 to 1200 where N is 300) is divided by N and multiplied by 100 to give the *preference index* As indicated, this value (with limits of 0 to 400) indicates the tendency of raters to mark people high or low on the particular behavior item As here computed, low values of the index indicate a tendency to mark the item as applying to a high or outstanding degree, high values indicate little or no applicability for the item

2 *Discriminative Index* For each of the alternatives of a given item, the difference between its frequencies in the upper and lower groups is computed These five differences are then added *without regard to sign* to give the *discriminative index* At one extreme, where the distribution of alternative frequencies is identical for the upper and lower groups, this index will be zero At the other extreme where the two frequency distributions have no overlap, the value will (in this case) be 200 Low values of the index, obviously, indicate that the item is equally applicable to good and poor officers and consequently does not discriminate High values, on the other hand, indicate gross differences between the groups in applicability of the item, and suggest that it represents behavior which has significance for success (or failure)

[The following chart] illustrates the method of calculating these two indices for a typical item

Item pairs are made up by selecting insofar as possible two items equal in preference value and widely different in discriminative value This selection is facilitated by plotting each item (identified by its number on the list) on a double entry table with preference values along the abscissa and discriminative values along the ordinate—both in suitable intervals By entering any row in this table, two items close in preference can be picked that are widely separated on the ordinate scale It is wise to avoid choosing items which are opposites in meaning since this eliminates the forced choice element Also, though the same item may be used in several pairs, it is wise to avoid too much repetition of this sort, since it tends to reduce the scope of the scale and necessarily raises the item intercorrelation, it may also inject an extraneous factor if the rater strives for consistency'

Finally, pairs of items are combined to form tetrads One pair with low preference indices (favorable) is combined with a second pair having high (unfavorable)

Setup for Determining Preference and Discrimination indices of Forced Choice Items

Alternative Weight (w)	1 0	2 1	3 2	4 3	5 4	
Frequency (f)						
Upper (U)	1	0	6	6	87	(N = 100)
Middle (M)	3	5	13	15	64	(N = 100)
Lower (L)	4	11	27	23	35	(N = 100)
Σf	8	16	46	44	186	(Σn = 300)
$\Sigma f w$	0	16	92	132	744	(Σfw = 984)
d|U − L|	3	11	21	17	52	(Σd = 104)

Preference index $\frac{\Sigma fw}{\Sigma n} \times 100 = \frac{984 \times 100}{100} = 328$

Discriminative index $\Sigma d = 104$

preference indices There is no logical basis for this step, but experience has demonstrated that if single pairs are used with instructions to indicate the most characteristic, there is considerable rater resistance to those pairs that have high (unfavorable) preference indices By combining high and low preference pairs with instructions to choose the most and the least characteristic, rater resistance is materially reduced The same end can be achieved by presenting high and low preference pairs (as pairs) separately with appropriate instructions for each

Scoring Forced Choice Rating Scales

Tetrads are formed from two pairs of items The members of each pair are matched for preference value One member of each pair differentiates good from poor officers The other does not It is possible, because of the way items are thus combined into tetrads, to key forced choice scales by assigning a point (plus or minus as the direction of the discrimination indicates) to each of the two discriminating members of each tetrad

Forced choice items may, however, act differently in combination with other items than they do by themselves Consequently, it is always desirable to establish the key on the final set of tetrads In doing this it is necessary to include enough tetrads so that those which fail to stand up on the final cross validation can be eliminated from the scoring key Needless to say, the cross validation should employ an external criterion

In one experiment on a group of 24 tetrads (96 items each of which could be marked as most or as least characteristic of the ratee), 75 per cent of the items were scored in the same way after cross validation as they would have been scored by a predetermined key based on the original preference and discrimination values It should be noted that while items which had discrimination value (either positive or negative) in the predetermined key may have lost their value, and while some items which did not discriminate in the original study came to do so in the cross validation run, there was not an instance in which an item which discriminated in one direction in the first experimental situation reversed its direction of discrimination in the cross validation run

The establishment of keys on the basis of a cross validation experiment rather than from the use of the discrimination indices increases the validity of the rating The experience of the Personnel Research Section indicates that the extra work involved in this additional step is justified by the increased validity that results from it

REFERENCES

1 Herron C D, Maj Gen, Efficiency Reports *Infantry Journal* April 1944 pp 30–32

2 Staff, Personnel Research Section The Forced Choice Technique and Rating Scales, *The American Psychologist* 1946 Vol 1 p 267

3 Witsell, E F, Maj Gen, The New Officer Efficiency Report *The Reser e Officer* 1947, Vol 24, No 6 pp 8–10

Measuring Supervisory Ability—A Case Study *

We are indebted to Matthew Radom of Jersey Standard s employee relations department and Edwin R Henry of Richardson, Bellows Henry and Company for supplying the factual material on which this study is based and for making many helpful suggestions regarding the analysis

PERFORMANCE RATING PROGRAM

INTRODUCTION

One of the most recent innovations in personnel administration is the use of psychological tests for selecting and developing supervisors It is only since the end of World War II that business concerns in significant numbers have begun utilizing tests and other systematic appraisal devices for the purpose of gauging supervisory ability The majority of these companies have adopted ready made tests, and have used the results without questioning or checking on the authors claims concerning validity A few, however, have undertaken internal research programs to develop measuring instruments geared to their own needs and to determine how well these instruments would actually assess the traits and skills required for effective supervision in the particular company

An outstanding example of the latter approach is the experiment in supervisor measurement conducted by the Standard Oil Company (New Jersey) This project was begun more than two years ago and is now nearly completed Its chief distinction lies in the fact that the principles and methods of social science research have been closely and consistently followed throughout

Two separate personnel appraisal instruments resulted from this investigation a procedure for rating supervisor performance and a battery of tests for selecting potential supervisors Careful validation checks have shown that both these devices are considerably more dependable than the ability measurement schemes commonly used in industry

A review of the Jersey Standard project and its results will, it is believed, be of interest to other companies confronted with the problem of achieving better supervision The present memorandum describes the development of the performance rating program and gives a brief appraisal of its value The construction of the test battery will be explained and its merits and limitations discussed in a second memorandum to be issued shortly

SUMMARY OF STEPS IN DEVELOPING THE PERFORMANCE RATING PROGRAM

The work of developing the Jersey Standard supervisory rating program and applying it on a trial basis was carried out in a series of separate but closely interrelated stages Before proceeding to the analysis of each stage in detail it will be worth while to get a picture of the investigation as a whole The steps involved in constructing and testing out the new rating scheme, briefly, were as follows

1 Meetings were held with the supervisory personnel of the plant chosen for the basic investigation in order to familiarize them with the purpose and nature of the project and enlist their co operation

* Reprinted from *Industrial Relations Memos* No 119, September 22 1950

2 The lower level supervisory force was ranked by the higher level management people to identify three known perform ance" groups of supervisors—high, middle and low, and a ' standard of reference" set of performance measurements was estab lished on the basis of the rankings

3 A comprehensive series of statements was prepared describing the various aspects of supervisory job performance in the plant

4 The statements were tested out by having upper level management men clas sify each statement in terms of how well it applied to the known performance super-visors identified in step 2

5 The results of step 4 were analyzed to determine with respect to each state-ment (a) the willingness of senior super-visors to use it in rating subordinate supervisors, and (b) the extent to which it differentiated between high performance and low performance supervisors

6 The experimental performance re porting instrument was devised by arrang ing the analyzed statements in ' forced choice' groups designed to enable senior supervisors to rate their subordinates care fully and impartially Two alternative re porting forms were constructed in this way

7 The known performance supervisors were rated by their superiors on the ex perimental report forms

8 The results of step 7 were compared, by careful statistical methods, with the standard of reference measurements estab lished in step 2

The statistical comparison revealed that approximately three fourths of the super visors in each of the three known-perform ance groups were correctly identified by the ratings obtained with the experimental forms As stated earlier, this is substantially more accurate performance appraisal than has resulted from any scheme previously developed for rating supervisors

Importance of Fair and Accurate Performance Rating

Prior to starting its development project the Jersey Standard management con-ducted an extensive survey of current methods and practices in the field of per sonnel measurement In canvassing various professional consultants, the management found that the great majority recom mended the use of ready made tests and forms based on their own concepts of good performance and constructive employee management relations

Nevertheless, the Jersey management did not feel that this approach was the answer to the problem of appraising employees more effectively To them it seemed in consistent with the wide intercompany differences in standards of performance found in actual practice They concluded that an appraisal instrument built wholly in terms of the situation in which it was to be used held better possibilities of fruit ful results

In line with this belief, they engaged the services of Richardson, Bellows, Henry and Company—one of the few firms con tacted which emphasized the individual company approach—to guide the develop ment program The technique employed by this group consisted in conducting on the ground research to define the existing standards in terms of actual employee per formance, and in devising tests and other appraisal instruments on the basis of these standards A member of the consulting firm and a staff specialist in the company's employee relations department were desig nated to carry out the research work

The location chosen for the basic inves tigation was the Baton Rouge refinery of the Esso Standard Oil Company, one of Jersey s principal operating affiliates The plant had recently been expanded, and a number of newly created supervisory posi tions had to be filled Moreover, the local management had felt for some time that its procedure in selecting foremen was not producing the best possible results, and consequently it was receptive to the idea of developing a more effective one The fact that the plant is a large one with super visory employees numbering in the hun dreds was a further advantage, since it enabled the investigators to make more meaningful statistical measurements than would have been possible in a smaller op erating unit

The research team, in consultation with

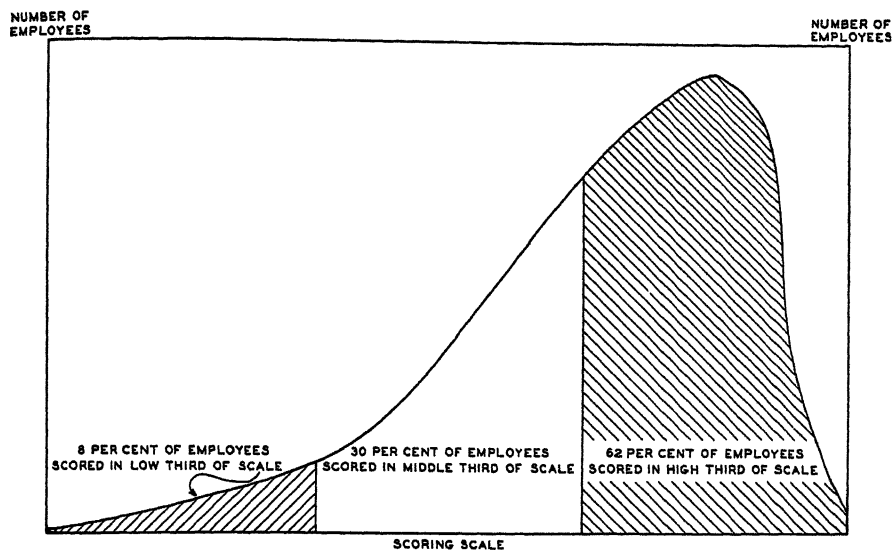



FIGURE 58 1 *Distribution of performance scores obtained on graphic scale report form of Standard Oil Company (New Jersey)*

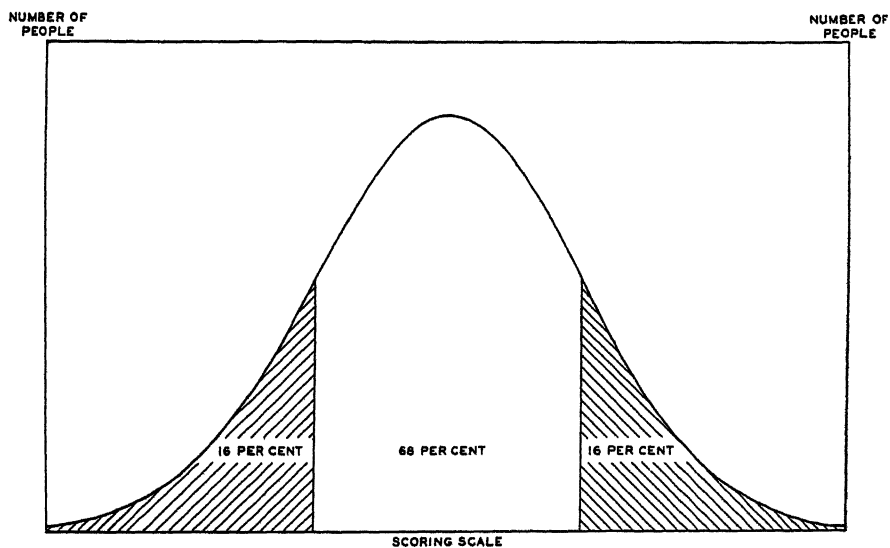



FIGURE 58 2 *Normal' distribution*

the refinery management, decided that the development of a genuinely fair and objective method of appraising supervisor performance must be the first order of business Such a performance rating scheme would be of help in isolating the abilities and behavior characteristics which distinguish successful foremen from less successful ones, and this in turn would aid in devising tests for selecting potential foremen of high caliber More specifically, it would provide a basis for checking the validity of any testing program which might be developed In addition, a consistent and valid performance reporting system in itself would be useful in selecting and promoting men from temporary to permanent foremanships and from the latter to higher levels Finally, it would be valuable as a counseling adjunct, in helping incumbent foremen improve their performance

The performance report then in effect at the Baton Rouge refinery and elsewhere in the Jersey system was of the graphic scale type—that is, a list of abilities and qualities relating to work performance, with a graduated scale opposite each item The scales were subdivided by descriptive phrases denoting degrees of progress or retrogression since the previous report, and with these statements as a guide each rater indicated his estimate of the subject

Although the graphic scale is used more widely than any other rating technique, it has two serious drawbacks One of these is that it usually results in a concentration of ratings near the upper end of the scale and a sparsity of ratings near the lower end This is attributable to the common tendency among raters to be overly favorable in recorded judgments of their subordinates, even of those whose performance is below par That the experience under the old Jersey report was no exception to this rule is apparent from Figure 58 1, which summarizes the ratings obtained over a period of years

This lopsided distribution of ratings was at variance with the known distribution of ability and work conduct among people generally, as measured by impartial observers trained in social and industrial psychology Typically, about two thirds of the individuals in a particular group fall in the middle or average capability portion of the performance scale About one sixth perform at higher levels than this, and the remaining sixth at lower levels This is the so called normal or bell shaped distribution, illustrated in Figure 58 2 Clearly, the distribution of ratings obtained on the graphic scale report diverged widely from the normal pattern

The other main defect of the graphic scale method is that it often results in biased ratings A supervisor whose judgment of his subordinates is colored by favoritism or prejudice will have a tendency to rate them unfairly high or unfairly low, unless there is some specific deterrent to his doing so The graphic scale provides no such deterrent, on the contrary, it may even accentuate the rater's propensity to include unconscious bias in his judgment because of the overlapping that usually exists between adjacent degree definitions on the scale

It was apparent to the investigators that the new performance rating procedure must avoid the defects of the scale then in use In other words, the procedure must result in an impartial and reasonably accurate appraisal of the degree of proficiency of the particular foreman in each significant aspect of his job In addition, it must be consistent in the sense that, if it were used to rate a man twice in succession, it would yield essentially the same results the second time as the first It must also be economical of time, in order not to place an undue burden on the reporting supervisors, and, finally, it must be useful as a counseling and training tool, to assist foremen in correcting their weak points and improving their effectiveness

Essentials of Forced Choice Rating

It was decided that the experimental rating program to be developed and tried out at Baton Rouge would be based on the forced choice technique In the opinion of the consultant, this method of judging performance offered the best possibilities of avoiding bias and overrating, while at

the same time meeting the other require ments[1]

Before going into the actual develop ment work, the basic features of the forced choice method will be explained briefly Boiled down to its essence it consists in presenting to the rater four alternative statements pertaining to job proficiency or relevant personal qualifications and asking him to indicate which one he believes is most applicable to the person he is rating and which one is least applicable Two of these statements are favorable in import and two unfavorable, as for example

Observes company rules
Handles people well

Afraid to make decisions
Does too much of the work himself

One of the favorable statements de scribes behavior or ability found only among high performance employees in the group being rated, while the other refers to an attribute found just as often among low- and medium performance as among high performance men (For brevity these may be termed, respectively, differentiat- ing and nondifferentiating' favorable statements) Similarly, one of the unfavor able statements is characteristic only of low performance men (differentiating un favorable), while the other applies to employees at all proficiency levels (non differentiating unfavorable')

This process of choosing alternatives is repeated with additional sets of statements until all the significant aspects of the job under observation have been covered Checking a differentiating favorable state ment as most applicable to a man, or a differentiating unfavorable statement as least like him, gives him a plus credit in the scoring Conversely, he gets a minus credit if a differentiating unfavorable statement is checked most or a differen tiating favorable statement least" All nondifferentiating statements are assigned zero credit, whether they are checked most or least An essential point here is that the key —i e, the knowledge of which statements are differentiators and which are not—is not available to the rater Only the person who scores the reports, usually a member of the personnel depart- ment, has access to this information Con- sequently the rater must make his most' and least' decisions without knowing whether they will have a determining ef fect on the ratee s standing or simply leave it in the middle"

One effect of this is that it removes the temptation on the part of the rater to shade his appraisal of a man out of per sonal liking or antipathy and consequently leads him to concentrate his attention on the man s actual performance in deciding whether or not he possesses the specified qualities In other words, it reduces the element of emotional bias to a minimum It also reduces the more pervasive tend ency to overrate medium- and low per formance employees, since the rater is obliged to consider whether the ratee has various stated shortcomings pertaining to the job and since, moreover, he does not know which ones will count against the subject and which ones will not

ESTABLISHING THE CRITERION

The foregoing merely summarizes the mechanics of the forced choice technique To understand the logic behind a forced choice rating system and arrive at an in- formed opinion of its value, it will be help- ful to know how the new performance report for supervisory employees was de- veloped and tried out at the refinery

In keeping with the general approach adopted the development and tryout work was conducted entirely in terms of the policies and standards of excellence in ef- fect in the refinery and of the supervisors actually employed there "Custom tailor- ing" is, indeed, a central feature of the forced choice method as it has been ap-

[1] Forced choice is a comparatively recent innovation in psychometric techniques It was first applied to the problem of apprais- ing merit in 1946, when a group of psy chologists in the United States Adjutant General s Office developed a forced choice performance report for rating Army officers See E D Sisson Forced Choice—The New Army Rating *Personnel Psychology,* Vol 1, No 3, Autumn, 1948, pp 365–381

plied in the field of performance rating The rationale of this approach and some of its implications will be discussed in a subsequent section

Prior to beginning any actual on the job research, the investigators held a series of meetings with all the supervisory employees in the refinery In these meetings the supervisors were given a detailed explanation of the purposes of the project and of the steps to be taken This would have been advisable in any case, but it was doubly so in this instance because the supervisors were to participate actively in each stage of the development process

The first task to be undertaken was the establishment of a "criterion, i e, a set of basic performance measurements to be used as a standard of reference for determining the accuracy of the experimental rating scheme This step consisted essentially in identifying a number of foremen of definitely known performance and, through a careful evaluation procedure, arriving at a numerical score for each man that could be termed a valid measure of his all around value to the company

The criterion was established by a procedure known as 'consensus ranking In each of the major divisions of the plant a number of lists of supervisors (mainly foremen) was prepared The lists were of two kinds—one including only men at the same level of responsibility and the other including men at various levels Care was taken to make certain that the performance of each man on a list was well known to several higher level supervisors Each of these senior supervisors, independently of his colleagues, was then asked to rank the men on the list identified with him, in the order of their overall value as practicing foremen The procedure followed was to name the most competent man, then the least competent, then the next most competent, etc, until the list was exhausted The number of senior supervisors ranking a given list ranged from 2 to 20 By this process the investigators obtained rankings of 32 lists comprising a total of 492 junior supervisors—an average of 15 individuals per list The rankings were then converted into 'standard' scores by statistical procedures to establish a valid common basis of comparison among the various lists

The several rankers of each particular list were generally in close agreement on the relative competence of the men listed However, in order to have a thoroughly dependable foundation on which to base the subsequent research, it was desirable to identify distinct categories of performance This was done by selecting three groups—high, middle and low—composed of foremen concerning whose standing the several rankers were in practically complete agreement To be included in the high group, a foreman must have had virtually all his rankings in the top 15 per cent of his group s score range Similar requirements involving the middle and lowest 15 per cent score ranges were established for the middle and low groups A total of 241 foremen was selected in this way The remainder were eliminated from consideration

Thus, when the process was completed, the investigators had identified three groups of foremen, each comprising about 80 individuals, which were clearly identifiable as high, medium and low performance groups in line with the practically unanimous consensus of senior supervision, who, in many instances, included top officials of the plant

Obtaining the Job Behavior Statements

The next step in the process was the collection and classification of a comprehensive series of 'job proficiency and 'personal qualification' statements A small but representative group of supervisors of all levels was designated as a behavior statement committee' Each member of this group was asked to pick out, but not name, an above average and a below average foreman with whose work performance he was thoroughly familiar, and to write a detailed description of each man and his actual behavior on the job Nothing pertinent to the manner of performing the work was to be omitted

These descriptive essays were then broken down into individual descriptive statements More than 1,000 separate behavior items" were obtained in this way

These items were classified into five cate gories relating respectively to manner of work performance, potentiality, skill in handling people, relevant personality traits, and executive abilities

Next, the entire collection of statements was submitted to a larger group of senior supervisors (in the main, the participants in the ranking procedure), who were asked to relate each statement to each of the foremen comprising the three criterion groups by use of a weight designation or symbol, as follows

Degree of Applicability	*Weight*
Fits the man exactly	5
Fits the man well	4
Statement and its opposite about equally true of the man	3
A rather poor description of the man	2
Does not fit the man at all	1
Statement describes something not applicable to the duties in the man s job	NA
Do not know the man well enough to make a judgment	U

Analysis of the responses resulting from this rating tryout revealed wide differences between statements in terms of the willing ness of the group of raters as a whole to apply them to their subordinates generally Some statements were frequently checked as highly applicable to foremen of all three proficiency levels, while others were as frequently marked not fitting at all at any level As one would expect, the for mer were in the main favorable sounding statements, and the latter unfavorable sounding On the basis of these differences the investigators were able to compute a preference index for each statement as a measure of its relative favorability—i e, of the average rater's relative willingness to use it The high index values denoted fa vorable statements and the low values un favorable ones Neutral statements were, of course, represented by the middle range of values

The analysis also revealed that certain favorable statements were nearly always labeled as fitting 'exactly" or "well' for members of the high criterion group, but seldom for those in the low group, and, conversely, that certain unfavorable state ments were termed applicable far more frequently to low performance than to high performance men As has been noted, still other statements—both favorable and unfavorable—were said to be applicable (or inapplicable) more or less uniformly to foremen of all three levels of profi ciency This variation in the differentiat ing power of statements made it possible to calculate indexes of differentiation ranging from zero for statements which were applied equally often to high and low men up to 25 for statements which were confined exclusively to high (or ex clusively to low) men

Constructing the Performance Report

The researchers were now ready to con struct the performance report form The first step consisted in compiling blocks of 'most least statements The general selec tion formula described earlier was fol lowed, except that five statements were included in a block instead of four Thus each block consisted of (1) a favorable statement that had been found to differ entiate between high and low proficiency foremen, (2) a favorable statement that did not differentiate, (3) a differentiating unfavorable statement, (4) a nondifferen tiating unfavorable statement, and (5) a neutral statement intended to apply to al most anybody in the group

In making the selection, care was taken to insure that the two favorable statements had approximately equal preference in dexes, and the same rule was observed in selecting the two unfavorable statements Another selection rule was that no two statements within a block should cover the same or closely related aspects of job per- formance And finally, the choice was care- fully spread among the several categories of statements so that, when all the blocks were completed, the various significant

EXHIBIT A

PERFORMANCE REPORT SUMMARY

This section is for summarizing the description indicated by choices on the preceding pages It should be completed in duplicate and one copy will be returned to the reporting supervisor for counselling purposes when desired Read each statement carefully and decide how well it describes the man To the right of each statement circle

1 if he is SOMEWHAT DEFICIENT the opposite or almost the opposite of what the statement indicates

2 if he is SATISFACTORY but not quite as good as the statement indicates

3 if he is EXCELLENT as good as the statement indicates

4 - if he is OUTSTANDING (Very few will be)

A He knows his job and his responsibilities has good qualifications and experience	4	3	2	1
B He thinks straight has sound judgment makes good decisions	4	3	2	1
C He has the push to tackle his work and can be depended on to see that it is carried through to a finished job in a workman like manner	4	3	2	1
D He plans organizes and delegates his work well	4	3	2	1
E He initiates and carries through improvements in methods of doing a job	4	3	2	1
F He practices good human relations is tactful and knows how to lead people	4	3	2	1
G He is effective in getting his ideas across	4	3	2	1
H He develops his men maintains good morale and discipline	4	3	2	1

Considering all employees that you know at his level of responsibility where would you place him in terms of over all value to the company ? (Check one)

☐ HIGH THIRD
☐ MIDDLE THIRD
☐ LOW THIRD

REMARKS

NAME OF PERSON REVIEWED ____________________ DATE ____________

REPORTING SUPERVISOR S SIGNATURE ____________________

DO NOT WRITE BELOW THIS LINE

FORCED CHOICE SCORE ________ SUMMARY SCORE ________ (SEE MANUAL FOR INTERPRETATION OF SCORES)

S-5

FIGURE 58 3 *The performance report form*

aspects of a foreman s job had been cov ered

The following is a fairly typical block resulting from this process

Most	*Least*	
A	A	Seldom makes mistakes
B	B	Respected by subordinates
C	C	Fails to follow through assignments completely
D	D	Feels his job is more important than others jobs
E	E	Does not express own views with any degree of self reliance

It is apparent that the two favorable statements, A and B, sound approximately equally complimentary It is not so ap parent, however, that item B is a good differentiator of high performance foremen—as determined by actual tryout at the refinery—whereas item A is not Hence the rater, in attempting to decide which of these items to check as 'most' (or 'least") for a particular man, has a real incentive to consider the man s actual per formance And similarly as between the two unfavorable statements C and E, since he is unaware that item C is more indica tive of low performance than E, he is again motivated to consider the man's real merits and shortcomings

After the compiling process was com pleted, the blocks were divided equally and assembled into two separate forced choice performance report forms which were designated as Form 'S" and Form "O," respectively In the final version, after some deletions, each form comprised thirty blocks In allocating the blocks, an effort was made to keep the job aspect coverage of the two collections as nearly the same as possible, with a view to obtaining forms that could be used interchangeably

A separate scoring system was devised for each form The method adopted for assigning score values to the statements within a block was essentially the same as the one outlined above except that low to-high positive values were used in place of a minus to plus range This permitted the use of simple addition in computing the total report score Since the maximum possible score range of Form 'S was greater than that of Form "O,' it was necessary to provide for converting the report scores into terms of a common or standard scoring scale

In addition to the 30 forced choice blocks, each report form as finally drawn up included a summary section This section was expressly designed for use as a counseling and training tool Since the relative rating significance of the various forced choice statements was not to be revealed, some such device was obviously needed as an aid to the reporting super visors in advising their subordinates and helping them to improve their performance The summary section, which was identical for both forms, is reproduced on the following page

The performance statements comprising this form summarize a substantial portion of the differentiating favorable" items in the forced choice form, but put in such general terms that they afford little possibility of ferreting out the key to the latter It will be noted that the summary report is of the same general type as the conven tional graphic scale report The investigators were fully cognizant of the limitations of this method However, its application here carried with it two qualifications which they felt tended to offset these weaknesses One of these was the explicit stipulation that the summary section was to be used for counseling and training only Consequently, it would not enter into the determination of the ratee s eligibility for promotion The other qualification was that the summary report should be made out immediately after completion of the forced choice section Experience with this type of form in other situations had shown that when this sequence was followed the propensity of raters to overrate medium- and low performance men was considerably reduced and that, consequently, the summary report appraisals were generally in fairly close agreement with the corre sponding forced choice scores In the few cases where marked disagreement was

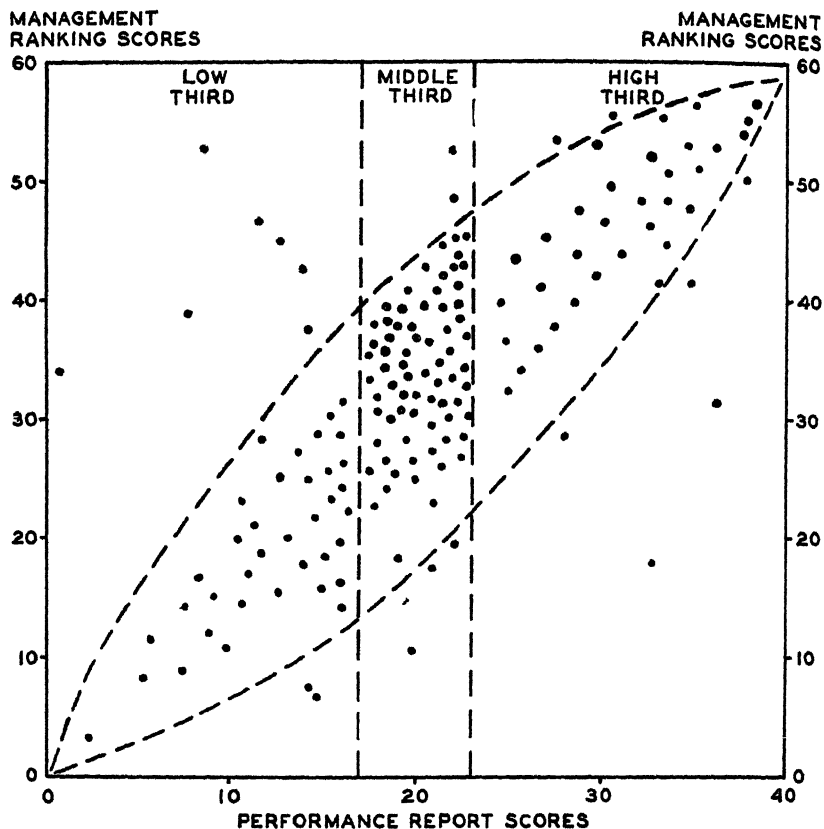



FIGURE 58 4 *Comparison of forced choice performance report scores and management ranking scores on foremen Baton Rouge refinery Esso Standard Oil Company*

evidenced, discussion with the rater had usually revealed the cause and suggested the remedy

Trying Out the Report

One more task remained to be done—testing out the performance report under conditions of actual practice to determine how well it would work The first requirement of a good performance rating procedure, it will be recalled, is that it should afford an accurate measurement of the ratee's proficiency in each aspect of his job The extent to which a rating procedure does this is termed its *validity* It should also yield the same results when used to rate a man twice in succession This quality is termed *reliability* [2]

The validity check was made by having report forms completed on each of the foremen in the three criterion groups in the same way as would be done under normal administrative conditions The first rating was made out by the particular foreman's immediate superior, using Form 'S' A second appraisal was made by another senior supervisor familiar with the foreman's work, in most instances using Form "O The resulting scores were then compared with the criterion—the array of

[2] Strictly speaking, reliability is a component quality of overall validity, since the term 'accurate measurement" implies exact sameness of results in successive trials However for present purposes the two terms may be treated as distinct without serious conceptual error

consensus ranking scores which had orig inally been established as the standard of reference

Figure 58 4 affords a good visual picture of the extent of agreement between the try out rating results and the criterion High performance report scores, clearly, are associated with high ranking scores, and low performance scores with low ranking scores to a marked degree Furthermore, the number of serious departures from this rule is only a small percentage of the correlation as used in statistics is defined as the relationship between two sets of variable characteristics or measurements, both of which pertain to the same group of entities—in this case a group of persons The correlation coefficient is a summary number which increases as the degree of relationship (for the group as a whole) increases, consequently it is useful as a convenient measuring gauge for judging the amount of relationship in a particular situation, and more especially for compar

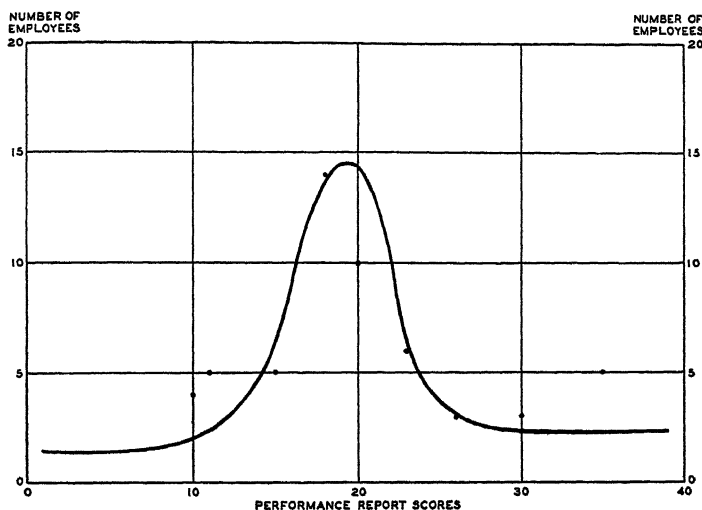



FIGURE 58 5 *Distribution of forced choice performance report scores on foremen Baton Rouge refinery Esso Standard Oil Company*

total It should be noted also that the performance report scores closely approximate the normal frequency distribution In Figure 58 5, the location of the dot above each point in the scoring scale represents the number of foremen who re ceived that score It is evident that this distribution bears a much closer resem blance to the normal curve than the distri bution of scores under the old graphic scale report (Figure 58 1) Thus the tryout re sults indicate that the use of the forced choice method does in fact greatly reduce the tendency to overrate medium and low performance men

A more concise way of measuring the degree of agreement is to use the Pear sonian coefficient of correlation The term

ing the degrees of relationship as between two or more situations

Although the computation of the cor relation coefficient is a rather involved statistical procedure, its essential meaning is fairly simple A coefficient of 1 00 indi cates a perfect (positive) relationship be tween the two variable characteristics, while a zero coefficient denotes no rela tionship at all To illustrate in the com parison under discussion, if the scores in the performance report tryout had come out in exactly the same order—man for man—as the scores in the consensus rank ing evaluation, the coefficient would have been 1 00 On the other hand, if the per formance report scores had turned out to be scattered in a completely random man

ner over the entire group—with some high ranking foremen receiving high scores and others low scores, and similarly for low ranking foremen—the coefficient would have been approximately zero

The correlation coefficients actually obtained from the comparison of the performance scores of the three groups of 'criterion' foremen with their consensus ranking scores are shown in Table 58 1 The other summary numbers, labeled standard error,' were also obtained by statistical computation The standard error gives an indication of how much the correlation coefficient would be likely to vary (up or down) if the entire experiment were repeated on a different but similarly constituted group of foremen

The results of validity checks on two other performance rating methods are available for comparison Various cases of performance measurement tryouts on foremen in which report forms of the graphic scale type were used yielded correlation coefficients of from zero to 0 30 Corresponding results in cases where standardized interviews were relied on gave coefficients ranging from 0 25 to 0 35

It is apparent from these figures that the S O forced choice performance report has a much higher validity than the rating reports based on the old fashioned graphic scale device, and that it is also greatly superior to the standardized interview method It is more difficult to appraise the validity of the forced choice report in absolute terms However, some idea can be gained from a hypothetical example of applying the forced choice results to the problem of selecting foremen for promotion Taking the case of the two report tryout, which yielded a coefficient of 0 74 if the highest third of the overall criterion group of foremen in point of performance report scores were selected, approximately three out of every four of them would be men who also fell within the highest third under the consensus ranking procedure, and the fourth would be a 'middle ranking man

It is true that this is still considerably short of perfect validity However, it should be borne in mind that the concept of validation as used here assumed that the criterion establishing procedure measured actual performance perfectly, and

TABLE 58 1

Correlation of Forced Choice Performance Report Results with Criterion Baton Rouge Refinery, Esso Standard Oil Company

Coverage and Application	*Number of Foremen*	*Co efficient of Correlation*	*Standard Error*
Entire plant, criterion groups			
Two reports on each man, different superiors, usually different forms (S' and O')	173	0 74	0 051
Form S one report	201	0 68	0 038
Form O one report	201	0 68	0 038
Maintenance and construction division			
Form S, one report	77	0 66	0 06
Production division			
Form S one report	75	0 65	0 07
Technical and engineering division			
Form S, one report	27	0 69	0 10
Accounting and employee relations divisions			
Form S," one report	22	0 77	0 09

that, consequently, any divergence from perfect validity must be attributable to the performance report Since the ranking procedure could hardly have been perfect, the true validity of the performance report is probably considerably higher than the coefficient of 0 74 would indicate

It will be noted also that the correlation coefficients for the separate plant divisions are nearly as high as those for the plant as a whole This means that the validity of the performance report as a tool for appraising supervisory capability is practically just as good within a particular division as it is when used on a plant wide basis

The 'S O" report was subsequently tried out in 9 other Jersey Standard affiliates It was found to have approximately the same validity for appraising supervisory personnel in each of these companies as in the Baton Rouge plant On 14 of the 15 groups of supervisors included in the tryouts, correlation coefficients (between performance report and ranking scores) ranging from 0 51 to 0 95 and averaging 0 67 were obtained The figure for the fifteenth group was 0 39

As a negative validity check the report was also tried out, in each of these companies, on the executive personnel ranking immediately above the supervisory category—namely, department and office unit heads The coefficients obtained on these groups ranged from 0 45 down to 0 01 with an average of 0 23—a conclusive indication that the S O forms were not well suited for rating managerial employees at the higher levels The most obvious reason for the low correlation is that these officials are concerned mainly with policy making and planning and that, consequently, direct dealings with subordinates constitute only a small part of their functions On the other hand the foremen and other lower level supervisors—on whose performance characteristics the reporting form was based—must be in constant personal contact with the men under their supervision In short, this test showed that the rating form was uniquely adapted to rating supervisory employees and that its uniqueness was a result of its having been developed in terms of the behavior of actual supervisors on the job In order to design a valid forced choice report for the higher category of personnel it would have been necessary to carry through a similar project based on their behavior

As a check on the reliability of the performance report, a sizable group of foremen were rerated by the same supervisor who had done the original rating but with the alternative form The resulting scores were virtually identical with those obtained in the original rating, indicating that the two forms can be used interchangeably and that each form is by itself highly reliable[3]

Merits and Limitations of Forced Choice Rating

Up to now, very few companies have adopted the forced choice method of rating, and fewer still have had more than a year of experience under it Consequently it is not yet possible to make a close appraisal of its value as an instrument for measuring supervisor performance It seems worth while, nevertheless, to consider briefly certain questions which have been raised concerning it, even though the discussion must of necessity be highly inconclusive

The questions are of two kinds—those relating to the process of developing the forced choice report and those having to do with its operating uses It will be convenient to consider these questions in the reverse order

Despite its newness, the superiority of the forced choice type of report over the older methods as an evaluative tool is already widely recognized[4] However, sev-

[3] The interform correlation coefficient was 0 93 This is substantially higher than the reliability correlations obtained on other performance rating schemes and equals those obtained on the best employment tests

[4] See, e g, E A Rundquist and R H Bittner, 'A Merit-Rating Procedure Developed by and for the Raters," in American Management Association, *Rating Employee and Supervisory Performance* New York, 1950, p 69, and D G Paterson Rating,' in D H Freyer and E R Henry eds, *Handbook of Applied Psychology,* New York Rinehart and Co, 1950, pp 149, 153

eral observers have questioned its effectiveness on other counts For example, one critic has stated that since the raters cannot be permitted to know the true values of the descriptive statements, the merits of this kind of rating are difficult to sell and explain to the raters' [5]

As a general rule, withholding from employees information which affects their jobs and work relationships is unwise It creates uncertainty, suspicion, and ultimately resentment In the matter of supervisors rating their subordinates, however, this factor is not always the main consideration Many conscientious raters find the process of rating under the conventional methods a burden, because they dislike the idea of deliberately giving below average men the low ratings which their level of competence warrants This is probably the main reason for the prevalence of overrating under the graphic scale method Such raters may actually prefer the forced choice method, since it simply requires them to consider and weigh specific facts concerning the ratee s performance, and leaves the evaluation of the facts to be done by persons farther removed from the scene, according to predetermined and impartial standards In other words, the secret key feature may be welcomed rather than resented

Moreover, the fact that the raters play a major role in the construction of the forced choice report form probably helps to influence their attitude toward it in the favorable direction At any rate the Jersey investigators up to now have found no evidence of reluctance on the part of senior supervisors to use the 'S O performance reports in rating their subordinates

Two other writers, in a recent article, have questioned the usefulness of the forced choice type of report for counseling and training purposes [6] These reports, in point of fact, are not directly usable for this purpose, owing to the necessity of keeping the behavior statement values secret As has been noted, the Jersey researchers were fully cognizant of this limitation and made specific provision for meeting it by stipulating that the Performance Report Summary be made out in addition to the forced choice form Since the summary is retained by the rater, he can use it as a guide in counseling with the ratee and helping him improve his performance

The point has also been made that when merit ratings are taken into account in promoting employees to higher positions and when the ratings are expressed in terms of a single numerical report score as in the forced choice report, inequitable decisions may result There is indeed a danger in placing too much reliance in the report score, since a small difference between the scores of two men in any given case may reflect inaccuracies of measurement rather than actual differences in performance The sponsors of the forced choice method recognize this danger and accordingly, recommend that report scores be left out of consideration in deciding promotions unless the score differences are substantially greater than the limits of accuracy of the reporting form

It is worth noting in this connection that Jersey Standard does not regard the S O report as being in any sense a short cut solution to the problem of selecting men from the lower supervisory ranks for promotion to more responsible posts The management is emphatic on the point that report scores should never be used by themselves in determining promotions The director of the personnel appraisal development research put the management's views on the matter as follows

We say that we are not trying to measure people so finely that a score of any kind will place one man above another man, but we do say that a performance report score should definitely place the man in the high, middle or low group of the particular population being measured—in this instance, all supervisors at a particular refinery or operation When we promote people from foreman jobs to higher levels of supervision, it is most likely that personal opinions around the committee table will still far outweigh any other considerations Undoubtedly if an individual s

[5] R E Shaeffer, 'Merit Rating as a Management Tool,' *Harvard Business Review* November, 1949, p 696

[6] E A Rundquist and R H Bittner, 'A Merit Rating Procedure Developed by and for the Raters," *op cit* p 69

performance report scores are consistently low, they will have the effect of minimizing his chances for promotion. On the other hand, if they are consistently high, they will greatly support his chances for promotion. If some of his report scores are high and some are low, this indicates a difference of opinion on the part of the raters. Thus, the decision will still have to rest entirely on the collective judgment of the committee or top management.

The question most often raised concerning the development process is one of cost. It is contended that the detailed on-the-job investigation and statistical analysis entailed in developing the rating scheme is so expensive that most companies cannot afford it. One commentator has stated that, owing to its costliness, the method can be used only by very large concerns.[7]

Undoubtedly some small firms would find it financially impossible to undertake such a program. The rating of supervisors, however, seldom presents a serious problem in small companies. It begins to be serious only when the supervisory staff becomes so large that top management is unable to observe individual supervisor performance at first hand. For companies of this size the one-time outlay involved in developing the program would not by itself debar consideration of the forced choice method. The question would be, rather, whether the long-term benefits to be gained from the measurement program were worth the cost of development.

There is, however, another factor relating to size of firm that must be considered. In order to construct and validate an effective supervisors' rating program within a particular company, it is necessary to have a supervisory force large enough to afford stable statistical measurements. That is to say, the "population" which forms the basis of the development procedure must be sufficiently large to yield sizable aggregations or clusters of measurements on each component group of men under observation. Otherwise the relative variation among the measurements may be so great as to preclude checking on the soundness of the development work or the validity of the rating program resulting from it. Generally speaking, a development project of the kind described in this memorandum would probably not yield dependable results unless the supervisory staff comprised 100 or more persons. Yet, even with this limitation, it is apparent that the applicability of the individual company approach is by no means limited to large concerns.

This leads us to another and more basic question—namely, whether the detailed development procedure is really necessary. Could not a "general" performance report form be devised that would accurately measure supervisory competence in all types and sizes of companies?

Undoubtedly certain personal characteristics and attainments relating to supervision are "universally desirable," in the sense that any company which employs supervisors possessing these qualities will benefit thereby. If all of these qualities were fully and precisely known, a forced choice form based entirely upon them no doubt could be devised. Such a form could be applied in all kinds of firms and situations without any individual company development work, and theoretically it should yield valid results wherever applied.

In the real industrial world, however, managements hold widely divergent views as to what constitutes effective supervision. It is primarily for this reason that the originators of the forced choice report insist that each program must be "custom built from the ground up." When the development work is completed in a particular company, the resultant set of behavior statements reflects, in essence, the attributes which *in the judgment of that company's management* distinguish good supervisors from poor ones, rather than a collection of intrinsically "good" or "desirable" characteristics. And when there is a wide difference of view on this score between companies, the valuations embodied in the rating systems developed will also differ—not merely in degree but in kind.

For illustration, compare two companies, one of which follows the policy of having

[7] R. E. Shaeffer, "Merit Rating as a Management Tool," *op. cit.*, p. 696.

all major decisions made at the top, and one which emphasizes decentralization of the decision making function In the former, unquestioning acceptance of orders by supervisors will likely be regarded by the management as an earmark of excellence, and in the latter, as an undesirable quality Consequently, if these two companies were to develop forced choice performance reports, a statement such as 'Never questions orders" would probably come out in the one case as a differentiating favorable item and in the other as a differentiating unfavorable item

With such contrasting views existing between managements, it would obviously be extremely difficult to design a general performance rating scheme, having a single set of evaluation standards, that would apply in all situations Nor would such a scheme be in any sense generally useful—assuming it could be designed If a management s conception of the essentials of good supervision is unsound or unrealistic, it would do little good to install a rating system based on a different conception, no matter how much sounder or more realistic the latter might be On the contrary it might do considerable harm, since the supervisors who received the highest ratings would probably be men who do not see eye to eye with the top management on basic questions of supervisory practice This could only lead to misunderstanding and conflict within the managerial group

Of course, little is to be gained in such a situation even from a program based on internal research It would simply result in measuring wrongly valued traits more accurately Thus, for example, in a company in which driving employees is considered more effective than leading them, such a program would have the effect of identifying and elevating more and better 'drivers"

In summary, then, the forced choice method provides a valuable appraisal tool when applied in terms of the policies and practices of the individual company, but only if the company s management is committed to a constructive and forward looking basic philosophy of employee relations

A Critical Review of the Validity and Rationale of the Forced-Choice Technique *

ROBERT M W TRAVERS

According to a paper by Staff, Personnel Research and Procedures Branch, the Adjutant General s Office (5), the basic idea for the forced choice technique was developed by Paul Horst with reference specifically to personality scales Robert Wherry developed a similar idea while working on personality measurement for the Civil Aeronautics Authority, and later the Staff of the Personnel Research Section of the Adjutant General's Office attempted to apply the concept first to the design of personality inventories and later to the problem of rating officers The latter application resulted in the production of the Army Efficiency Report, which was used for some years for the official rating of officers More recently the technique has been used by various professional persons and for a number of different purposes The method of making a forced choice scale as used by the Personnel Research Section has been well described by Sisson (3)

The general nature of a forced choice rating scale is best understood in terms of the procedure which is followed in building it The following are the essential steps involved

* Reprinted from the *Psychological Bulletin*, Vol 48, No 1, January 1951

1 Descriptions are collected of individuals who are at each extreme of the scale to be measured If the scale is that of effectiveness as a supervisor, descriptions are obtained of the most effective and the least effective supervisors This procedure partly defines the scale that is to be measured by the final instrument

2 The descriptions collected are then dissected into a list of small elements Each one of these elements describes, in essence, a rather specific item of behavior The complete list should cover all the important aspects of the job, and the number of items covering each aspect should be related in some rational way to the importance of that aspect

3 Two values are determined for each item listed One value, the discrimination value, indicates the degree to which the item measures the particular characteristic that is being measured The other value indicates the extent to which individuals tend to rate others too high or too low on the particular characteristic This latter is sometimes referred to as the preference index or preference value of the item Both values are determined experimentally

4 The characteristics are then arranged in pairs such that the two members of each pair differ in the extent to which they discriminate Ideally, one of the characteristics in each pair should have a discrimination value of zero and one should have a high discrimination value Also, it should be impossible for those who are to use the scale to determine by inspection which one of the two characteristics is the discriminating one

5 The pairs of characteristics may then be grouped in fours, with each group of four including two desirable and two undesirable characteristics The main purpose of grouping the characteristics in groups of four seems to be that persons filling out the scale may have a better attitude towards it if the grouping in tetrads is adopted This seems to be the only advantage of the tetrad over the duad form A pentad form has also been suggested

6 Directions are prepared in which the individual is instructed that he is to examine each group of four characteristics and to select the one that is most characteristic and the one that is least characteristic of the person who is being rated The person filling in the form is required to make these choices

7 The selection of the items is then validated against an external criterion on a sample that was not used in the original procedure for selecting and pairing the items

The scoring system is such that the person who is being rated receives a positive score if the item which is most descriptive of him is a discriminating desirable characteristic or if the item least descriptive of him is the undesirable discriminating item The scores on items may be weighted in some complex manner or the weights may be restricted to the values of zero and unity

What Does the Rater Do in Completing a Forced Choice Rating?

The task performed in rating another person on a forced choice scale is somewhat different from that performed in any other phase of life Consider, for example, the simplest situation in which two items are presented and the rater is asked to decide which one is most characteristic of the person rated, and suppose that the scale was to be used for evaluating some phase of administrative ability and that the following items constituted the choice offered

> Enjoys making speeches at company dinners
> Contributes to company magazine

Suppose that the rater has good evidence to show that the person being rated has a matter of fact attitude toward the task of making speeches at company dinners He does not get any particular enjoyment out of speech making, but neither does he find the job distasteful As far as contributing to the company magazine is concerned, he dutifully writes an article every year, as do most executives of the particular company The situation is confusing for the rater, to say the least, since he is required to compare an item of behavior which varies in degree with one which varies in frequency

Yet the confusion in this instance is much less than that produced by some of the items in the Army Officer Efficiency Re port In the latter connection, consider the following tetrad of items which appears in the Army scale (WD AGO Form 67 1 Part 2)

A go getter who *always* does a good job
Cool under *all* circumstances
Doesn't listen to suggestions
Drives instead of leads

In the example, the two words italicized were not in italics in the original version If these two words are carefully examined in content it will be found that their in clusion makes it impossible to compare these items with the others Quite ob viously a man either is or is not cool under *all* circumstances It is not an item of be havior which varies in degree Under such circumstances the rater is asked to perform an impossible task He cannot say that one of a set of items is more characteristic of the person being rated than any of the other items listed, because some of those listed either are or are not characteristic of the person being rated regardless of the other items in the tetrad

This confusing situation is mainly a re sult of the way in which forced choice scales have been constructed and is not a necessary consequence of the forced choice technique

How Can the Choice Be Made Rational?

It is anticipated that one of the neces sary changes in the method in order to develop its usefulness as a measuring de vice will be to modify it in some way so that the rater makes choices only in situa tions in which choice is rational One step in this direction would be to include in each tetrad only items of behavior which vary in frequency of occurrence within a given individual or only items which vary in degree It hardly seems possible to com pare behaviors which vary in frequency with behaviors which vary in degree

The second necessary step in clarifying the method of measurement is in defining what is meant by the most characteristic' or least characteristic Like many terms these seem clear until they are carefully examined for meaning For example, when a rater is asked to select from two items of behavior the one that is most character istic (or more descriptive, or whatever else), he is really being asked to choose the one on which the individual deviates most from the average or from some other stand ard If two items of behavior which vary in frequency are under consideration, a comparison of the frequency of one item of behavior must be made with the fre quency of occurrence of the other item of behavior in the person rated This com parison should presumably be made not in terms of absolute numerical frequencies but in terms of some converted score which takes into account the relative scat ter of scores of the two variables The rater might be asked to choose the character istic on which the individual would have the highest relative standing in his group (which would have to be defined) if all individuals were ranked in order If the directions could convey this kind of con cept it might help to bring meaning into a confused situation

However, having developed and clarified the psychological concept to this point, the reader will probably have already jumped ahead to ask the question Why is it necessary to force the rater to choose between the characteristics? Would it not be simpler to ask him to rate the individual on each one of the two characteristics and then examine his ratings to determine which one of the two characteristics was rated higher? The answers to these ques tions are in the affirmative, since, if the directions are clear, it is necessary for the rater to assess the individual rated on each one of the items of behavior listed before he can rationally determine which one is most characteristic or which one is most descriptive' The only additional restric tion placed by the forced choice method is that the rating on all the characteristics in a given group must be different, other wise the forced choice scoring system can not be used in its present form

If the rater is asked to make an assess ment of the individual to be rated on each

of four characteristics, but with the added restriction that no two ratings within any group of four may be placed at the same point on the scale, the resulting record may be scored in exactly the same way as any of the forced choice scales at present in use In other words, the process of forcing a choice is an unnecessary and irrelevant part of the procedure However, what is left when the procedure is rationalized in the way shown is a scale in which tendencies to overrate or underrate on relevant qualities can be corrected by overrating or underrating on irrelevant qualities It is assumed in this procedure that, for any given individual, the true rating on the irrelevant items is average and that any deviation in the average ratings on these irrelevant characteristics represents a tendency to overrate or underrate the particular individual who is being rated The relevant items can be selected empirically, in the first place, to make this assumption acceptable The procedure is simply that of using ratings on certain characteristics as a suppressor variable to correct for errors in the rating on certain other characteristics In this form the basic idea of a forced choice scale seems to have considerable merit and to be one worth developing though possibly in forms other than the stock in trade forced choice instrument at present in use

The Validity of the Forced Choice Rating Scale

The claims for the validity of the technique seem to bear little relationship to the actual evidence

A primary claim is that it produces a better distribution of ratings relatively free from the usual pile up at the top of the scale (3) The data in Figure 59 1 were presented to substantiate this claim

Figure 59 1 is supposed to compare the distribution of ratings of officers when the old conventional rating scale (Form 67) was used and when the new forced choice rating scale (Form 67 1) was substituted It is quite obvious to anyone who examines the above graph that differences in the two distributions do not at all substantiate the claim that the forced choice rating technique is relatively free from the usual pile up at the top of the scale This claim

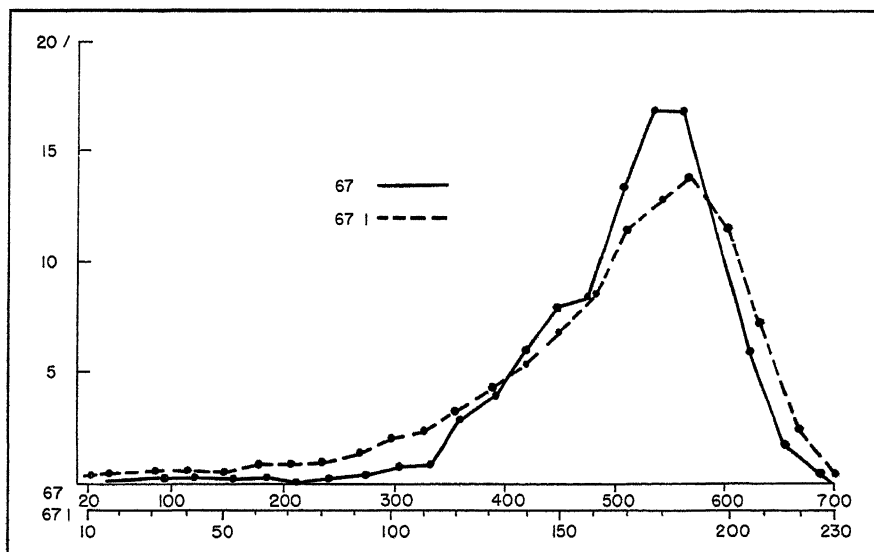



FIGURE 59 1 *Data presented by Sisson showing distribution of ratings on conventional rating scale Form 67 and distribution ratings on the forced choice rating scale Form 67–1 (3 p 375)*

is even more ridiculous when it is sub stantiated by Figure 59 2, which presents the contradiction of stating that it is based on data in which range is equated, in spection of the data seems to indicate that the ranges of the two sets of ratings were not equated (data from J C Fry)

A second claim is that the forced choice rating scale as used by the Army pro duces ratings which are more valid indices of real worth" (3) Similar unsubstantiated claims come from Richardson (2), but the apparently spurious nature of the validities presented by him and his group will be discussed later Claims contradictory to the above are also made by professional personnel who worked on developing the Army Report Form, as is exemplified by the following statement by Staff, Personnel Research and Procedures Branch, the Adjutant General s Office (4)

> A single over all rating on a 20 point scale, provided it is preceded by a series of specific ratings—using either the paired or unpaired items that have been de scribed—is more valid than either type of specific item alone

FIGURE 59 2 *Data presented by Fry showing distribution of ratings on conventional rating scale Form 67 and distribution of ratings on the forced choice rating scale Form 67-1 (1 p 22)*

Claims for superiority of the forced choice rating scale and claims for the su periority of traditional methods coming from the same research office lead one in evitably to conclude that available data cannot give strong support to the supe riority of the one method over the other in so far as Army studies are involved The following quotation summarizes all the

> statistical evidence that the present writer has been able to obtain concerning the relative validity of the forced choice (FCL) and the more traditional type of rating device (RCL) as it is used in the Army Officer Efficiency Report (4)
>
> The FCL key contains 48 items, the RCL key, 93 items On the basis of length of rating scale the RCL key would be ex pected to yield the highest criterion cor relations This is the case for the items scored within the OER-B form, although the difference for the larger group is slight (r s equal 502 for FCL score and 513 for RCL score) When the FCL form and the RCL form are given separately, however, its difference favors the FCL form (r's equal 466 for FCL and 345 for RCL) These findings suggest that when used alone the FCL technique is superior to the RCL technique That the two tech niques yield similar results when used in combination may well be owing to the forcing of more serious consideration of the individual items when it is known that a choice must be made after rating on each single trait
>
> Combining the two techniques yields the highest validity for both the item analyzed group (540) and the cross validation group (562) The gain is considerable when compared with the validity for the RCL technique (345) or for the FCL technique (466) used alone, but negligi ble when compared with the validity for the RCL score obtained when the tech

niques are used in combination (513 for the item analyzed group and 535 for the cross validation group)

The implication in the above quotation almost seems to be that the presence of forced choice items has the magical effect of raising the validity of traditional rating procedures Any such interpretation is obviously nonsense, for it was shown in the same series of studies that a single over all rating of effectiveness, of the traditional type, was more valid than the forced choice scale provided it was preceded by a detailed rating system which might be of either the traditional type or the forced choice type What the data do show is that the validity of a rating scale is likely to vary to a marked degree with the orientation given to the rater The data suggest that experimentation with different types of directions may yield much more important results than experimentation with forced choice scales

In whatever way these matters may be interpreted, the fact remains that the traditional type of detailed rating procedure and the forced choice procedure have approximately the same validity when orientation is given, and an over all rating scale of officer effectiveness may have higher validity than either of these procedures

Other Validity Studies

Other material presented as evidence of validity is supplied by Richardson (2) The present writer cannot accept the evidence provided by that author and his associates since the procedure involved seems to raise spuriously correlations between assessments of job performance based on a forced choice scale and an independent criterion of job proficiency The Richardson technique seems to result in spuriously high validities because of the technique used for refining the criterion The criterion of supervisory ability is obtained from judgments of those who are in a position to judge the supervisors Where judges disagree on the rating to be assigned to a supervisor, Richardson discards the case What is happening, of course, is that those supervisors who are at the extremes of the rating scale tend to be retained while those in the middle of the distribution will tend to be rejected, since those at the extremes can almost inevitably be more accurately judged than those in the middle range The effect of Richardson s procedure on the distribution of cases on the criterion measure must be considerable since he reports that 51 per cent of the cases were rejected It need hardly be pointed out that a correlation between the criterion and the forced choice scale is almost inevitably raised by increasing the relative proportion of cases at the extreme ends of the distribution particularly when the reliabilities of the criterion rating varied from only 0 4 to a (chance) one of 1 00 ' It should be noted in this connection that the validity coefficients supplied by Richardson apply to hypothetical populations of cases on which raters agree and do not apply to the population on which the scale was to be used Such data provide little evidence concerning the validity of the scale when used for the purpose for which it was designed

Another point to note in connection with the Richardson study is the use of the concept of reliability While the study shows that a rater is consistent with himself—both throughout one form of the forced choice scale and from one form to another—in the only case in which different raters used different forms of the rating sheet the reliability of the resulting rating was only 0 69 Also it is not clear whether this reliability was determined on the attenuated or on the original population before cases were discarded If this particular reliability coefficient, and the others for that matter, were calculated on the basis of the reduced population (with 51 per cent of all cases discarded), they would probably be numerically larger than if the entire population was used

Other Criticisms of Validity Studies

There is, however, a basic criticism which must be made of all the validity studies which have been reviewed, namely, that they all involve the unsatisfactory process of predicting ratings with ratings Studies which involve criteria other than judgments must be made and must form

the ultimate basis for evaluating the effectiveness of various assessment techniques

Another claim made by the protagonists of the forced choice technique is that it prevents the rater from controlling the value of the assigned rating It is fairly obvious, however, that this is not so If a rater decides he wants to give a person a higher rating for job proficiency than is deserved, all he has to do is to think of the person who was generally considered to be most proficient in that job and to fill out the forced choice form for that person The technique would almost certainly ensure a high score In a similar way a low score could be assigned at the will of the rater The method which a rater has to use to control the rating is a little more complex with the forced choice technique than with the conventional rating technique It is still an open question as to what fraction of raters will see this loophole in the technique, but loopholes of this kind are likely to become common knowledge

General Summary

A careful examination of the forced-choice technique as it is used in the Officer Efficiency Report reveals serious defects in the rationale which result in the rater having to make judgments which are strictly illogical These defects in the technique are not, however, basic and can be remedied

An examination of the validation studies of forced choice assessment methods reveals that the evidence does not support some of the claims made for the validity of these procedures The results show that in studies in which little criticism can be made of the procedure, there seems little to choose between forced choice and traditional rating The high validity coefficients secured by Richardson and his associates must be considered to be largely spurious until they are demonstrated to be otherwise

While validity data published at the time of this review must be considered extremely unsatisfactory and, on the basis of this evidence, the technique does not seem particularly promising Proper studies need to be made to determine the validity of scales in this area constructed on the basis of an adequate rationale

Bibliography

1 Fry, J C, 'All Superior Officers, *Infantry Journal* 1948 Vol 63 21-26

2 Richardson, M W, An Empirical Study of the Forced Choice Performance Report A paper presented at the 57th annual meeting of the American Psychological Association, Denver, Colorado, 1949 (Mimeographed) *American Psychologist* 1949, Vol 4, 278-279 (Abstract)

3 Sisson, E D, Forced Choice—the New Army Rating, *Personnel Psychology,* 1948, Vol 1, 365-381

4 Staff, Personnel Research and Procedures Branch, Adjutant General s Office *Comparison of the Rating Check List (RCL) and Forced Choice List (FCL) Methods of Obtaining Ratings* P R S Report No 717, 9 July 1946 (Mimeographed)

5 Staff Personnel Research and Procedures Branch, Adjutant General's Office The Forced Choice Technique and Rating Scales A paper presented at the 54th annual meeting of the American Psychological Association, Philadelphia, Pennsylvania, 1946 (Mimeographed) *American Psychologist* 1946, Vol 1 267 (Abstract)

Reply to Travers'[1] *"A Critical Review of the Validity and Rationale of the Forced-Choice Technique"* *

DONALD E BAIER

The Personnel Research Section of The Adjutant General s Office is an operating research agency It conducts both basic and applied research for the purpose of providing the Army with the best possible personnel tools As an operating research agency, it sometimes, because of the pressure of events, delays publication of results of interest to the psychological profession By the time research is published, all too often progress has been made toward further refinement or, sometimes, even toward a different viewpoint on a given problem The recent review by R M W Travers of the forced choice technique (23) is a case in point

Travers is in the unfortunate position of attempting a review of a problem which originated in the Personnel Research Section and on which but a small fraction of the Section s research has been published in the psychological journals However, reports of additional research are available as Personnel Research Section Reports[3] and as papers read at meetings of the American Psychological Association This additional information would have kept Travers from making such statements as the following (23)

1 The following quotation summarizes all the statistical evidence that the present writer has been able to obtain concerning the relative validity of the forced choice (FCL) and the more traditional type of rating device (RCL) as it is used in the Army Officer Efficiency Report (p 67)

2 The claims for the validity of the technique seem to bear little relationship to the actual evidence (p 66)

3 Proper studies need to be made to determine the validity of scales in this area constructed on the basis of an adequate rationale (p 70)

Travers could have obtained the additional information had he but indicated his intent to write a review The difficulties of generalizing in the field of personnel research resulting from the operation of so many factors, are well known A critical review is scarcely worth the name if it leans heavily on a single study and does not cover other research information

There is one highly important omission from Travers article His title and his remarks are directed at the forced choice technique in general, although the data he discusses are confined to rating procedures It should not be overlooked that the forced choice technique has been used in other types of instruments, for example, personality inventories and self description forms The forced choice technique has demonstrated its great usefulness in the construction of such instruments (12, 13, 16, 19) Even higher validities have been obtained with modifications of conventional forced choice technique based on suppressor theory (1, 20) The lack of explicitness in Travers' review on this point is misleading Compared to the traditional "yes no' type of questionnaire, the tech

---

* Reprinted from the *Psychological Bulletin* Vol 48 No 5, September, 1951

[1] Travers, R M W A Critical Review of the Validity and Rationale of the Forced Choice Technique *Psychological Bulletin* 1951, Vol 48 62–70

[2] The opinions expressed herein are those of the author and do not necessarily reflect the views of the Department of the Army

[3] Personnel Research Section Reports are not available for general distribution Arrangements have been made to furnish the American Documentation Institute with copies of unclassified reports for distribution

nique has produced personality measures of useful predictive value for the Army situation Working with the traditional type of personality items has consistently failed to yield a useful product

The comments in this reply refer only to rating scales used for efficiency report ing or merit rating purposes as involved in the work of the Personnel Research Sec tion[4] Further, the paper is not intended as a complete review of the research con cerning this application of forced choice It is an effort to discuss the problems raised by Travers, and to make available some of the more general findings concerning the problem of efficiency reporting This effort will be made in terms of (1) areas in which there is agreement with Travers, (2) areas in which there is disagreement with him, and (3) areas or problems concerning which he makes no comment

Areas of Agreement

Ratings should not be validated against other ratings (23, p 69) From one point of view there is no doubt of the desirability of criteria other than judgments The Personnel Research Section and other investigators have searched and still are searching for more objective and appro priate criteria Practicable suggestions for the development of such criteria would be eagerly welcomed Until such a develop ment occurs, investigators will no doubt continue to use ratings as criteria, and considerable effort will be expended to ward improving such ratings

The problem of criteria for efficiency reports deserves more than the brief com ment Travers gives it In considering the use of ratings versus objective criteria in any instance, the nature of success being predicted must be carefully considered In some instances, ratings may be the best criteria because value judgments are the essential elements What must be clearly recognized also are the problems of interpretation that are involved when ratings are used to "validate" ratings More specifically, three points should be noted concerning the use of a composite of ratings as a criterion

1 It is considerably better than no cri terion at all It is well known that averag ing a series of ratings will tend to reduce bias At the very least, therefore, use of multiple ratings as criteria in evaluating rating scales will improve rating proce dures by identifying those scales which con tain the least amount of bias

2 Use of composite ratings as a cri terion would seem to have its maximum justification in studies of performance as an officer This is the case, since perform ance as an officer involves, as a large com ponent, the ability to work with and through other people Furthermore, an of ficer s career involves a large variety of duty assignments, the expression of his value must be in generalized terms Judg ments of superiors, subordinates, and im mediate associates are especially pertinent

3 Use of multiple ratings as criteria creates problems of interpretation of the findings involving comparison of specific rating techniques Up to the present, the rating composite has essentially been an averaging of ratings obtained by a single technique—the traditional type of rating scale When a rating scale is involved as a predictor, one never knows the extent to which it is favored because of its similarity to the criterion Indirect evidence suggests that the amount of such 'technique con tamination is appreciable (17) The solu tion to the problem of comparing rating techniques, when criteria differing entirely in character are unavailable, may be the inclusion of all types of rating techniques in the criteria This procedure will give each rating technique an equal opportunity of showing validity" Such a procedure reduces the problem of 'validity' in rating studies to one of rater agreement, i e, re liability, if this concept is considered to cover a relationship of one rater using a given technique to several raters using the same technique

In the sense of choosing between two members of a pair, forcing a choice is not an essential part of the technique In dis cussing the rationale of the technique,

[4] The problems involved in securing ratings for criterion purposes are not necessarily the same

Travers makes a great deal of the point that all items of a pair or a tetrad could either be listed in rank order or the rating could be given in terms of a traditional rating scale with the restriction that no two traits could be rated at the same point (23, pp 64–65) This may well be true The possibility has already been indicated in conjunction with self rating items (7, p 186) Inasmuch as it is obviously difficult, if not impossible, to extend indefinitely the number of items which can be considered together, an element of forcing is bound to be present This is nothing new Choices must be made among the terms that are grouped together, much in the same sense that a choice must be made among the alternatives of any multiple choice item This point is relatively unimportant except for its relationship to the next

Forced choice pairs work because the nonscored alternative serves as a suppressor This is an important point because of its theoretical significance We agree that suppressor theory may provide the rationale for the success of forced choice items, in fact, we have exploited it heavily in connection with self description inventories, as mentioned above (20) To avoid any misunderstanding, however, certain points should be made explicit

1 Travers states that forced choice procedure assumes for any given individual, the true rating on the irrelevant [i e, unscored] items is average and that any deviation in the average ratings on these irrelevant characteristics represents a tendency to over rate or under rate the particular individual who is being rated The procedure is simply that of using ratings on certain characteristics as a suppressor variable to correct for errors in the rating on certain other characteristics" (23, p 65)[5]

2 The above assumption is not necessary, nor was it made in the development of the forced choice procedure To quote from one of the Section s early papers (11) on this technique, The essence of the forced choice technique, as we use the term, however, is the grouping of the alternatives to make them appear of equal value, and yet have unequal significance In other words, items are paired so as to give each alternative equal face validity and differing true validity Whether or not individuals have an average rating on irrelevant items is not essential for either of these conditions to obtain, nor has it any bearing on suppressor theory in forced-choice items

3 The suppressor theory requires only that (*a*) the scored alternative of a pair have as high a validity as possible, and (*b*) the nonscored alternatives have as low (even negative) a validity as is consistent with a high relationship with the scored item

4 A casual reader of Travers' article might conclude that a separate suppressor key could be developed for the traditional rating scale items Travers did not suggest this possibility, but it is desirable to make explicit the point that this application of suppressor theory may not work Use of traditional ratings as suppressors for other traditional type ratings has been tried in obtaining rating criteria The suppressor theory has been confirmed in the sense that negative Beta weights were obtained for the intended suppressor ratings, but the effect was so slight that validity of the combined ratings was not improved (18) It is not intended to assert that a suppressor key might not be developed on the basis of traditional type items, but only that available evidence does not encourage optimism in this belief

Forced choice items can be improved in their content (23, pp 63–64) Pairing items on a statistical basis only will frequently bring together alternatives which normally would not be associated Indeed, this may be one of the advantages of the forced choice procedure Travers' point that the content of the pairs should not confuse the rater is well taken However, we believe he has exaggerated the problem Directions to the rater have always

[5] 'Irrelevant and relevant" are perhaps not sufficiently meaningful in this context In one sense, all items are relevant" to officer performance 'Discriminative" or nondiscriminative,' 'differentiative" or nondifferentiative,' "critical' or "non critical are suggested substitutes

stressed that he is to indicate which alternative *most nearly* applies to the person he is rating Items such as Travers cites do meet the crucial test of having and *maintaining validity* over a period of time (8)

Forced choice items do not prevent the rater from manipulating his rating if he so desires (23, pp 69–70) In publications of the Personnel Research Section, claims have been much more modest than Travers implies To cite one instance, it reduces the rater's ability to produce any desired outcome of obviously good or obviously bad traits It, thus, diminishes the effect of favoritism and personal bias' (10) The emphasis is on the words "reduces' and diminishes"

Personal bias is a general term indicating departure from the true value for any reason Bias may result from insufficient information on which to base a rating, from the unconscious operation of friendship, from differences in leniency on the part of raters, etc It is in the reduction of these types of bias that the forced choice technique may be particularly helpful The rater who deliberately desires to manipulate his rating can undoubtedly do so However, the forced choice technique makes it a somewhat more difficult task for him

In passing, it might be pointed out that an efficiency report is primarily a means of recording the rater s estimates By itself, regardless of technique used, it does not guarantee that the rater will be honest, comprehensive, careful, and objective To achieve this purpose, supplementary aids must be used, and even these may not be effective In the Army, this aid is in the form of an Army Regulation which contains not only the necessary administrative procedures but also a discussion of the purpose and use of the efficiency report and of the psychological principles involved in rating This psychological information would not have been included in the Regulation if it were believed that the forced choice technique were an automatic and complete control of rater bias

In relation to this question, it should be pointed out that while a rater can move his rating up or down the traditional type rating scale at will, and can influence the score he is giving on forced choice items, on neither type of rating scale can he determine with much precision the relative standing of the person he is rating This point is most clearly seen when scores on rating scales are translated into some standard scale It is not uncommon on a 7 or 8 point rating scale for 30 per cent of the responses to be concentrated at a single point The amount of change a swing of one point on a scale will produce on a standard score is evident Unless, therefore, a rater knows precisely the distribution of ratings, he can never know where he is placing a person on a relative population scale, the kind of rating used by the Army This point is mentioned because it is believed that a good deal of the objection encountered by the forced choice technique has been misdirected, and the point to which objection is taken is basically the difficulty of reconciling relative and absolute standards

Areas of Disagreement

Some of the areas of disagreement are quite minor These will be disposed of first

1 *Travers is incorrect in his statement 'Each one of these elements describes, in essence a rather specific item of behavior'* (23, p 62) A glance at the alternatives shows very many general terms, i e, *modest no one ever doubts his ability, low efficiency businesslike* One of the unsolved problems is the degree of specificity which alternatives in forced choice groupings should possess

2 *Travers is incorrect in his interpretation of preference index* He states, 'The other value [preference index] indicates the extent to which individuals tend to rate others too high or too low on a particular characteristic (23, p 62) The preference index is, to quote from an early publication on this technique (11), an index of the "value to the rater' of the alternative under consideration, more recently, the preference index has been considered as a measure of the face validity of the item It is hoped that use of the forced choice technique will tend to correct for raters

tendency to rate too high or too low, but this is not involved in the computation of the preference index

We disagree with Travers' statement "*Claims for the validity of the technique seem to bear little relationship to the actual evidence*" (23, p 66) In support of this statement, Travers relies heavily on a minor study of the Personnel Research Section (14) and one by Richardson (6) As mentioned earlier, in making this statement he has ignored the vast body of research data available Some of the data have been presented at recent meetings of the American Psychological Association (2, 4, 8, 9, 21)

tained by a nominating technique The consistently greater validity of Form 67 1 is evident

We disagree with Travers' interpretation of the quotation "[the forced choice rating technique is] relatively free from the usual pile up at the top of the scale (23, p 66) In the first place, Travers has confused the forced choice technique per se with Form 67 1 This form contains both forced choice and traditional type rating scales The distributions he reproduces (from 10) are for the total score on Form 67 1 Travers does not observe this distinction, hence, his remarks are misdirected

In the second place, Travers does not comment on the difficulties in comparing

TABLE 60 1

Comparative Validity of Form 67 and Form 67 1, April, 1946, Edition (from 15)

Rank	*Sample 1* (*N* = 4,208)		*Sample 2* (*N* = 3,563)	
	Form 67	Form 67 1	Form 67	Form 67 1
Col	24	35	30	30
Lt Col	13	23	48	50
Maj	32	42	32	34
Capt	21	31	34	35
1st Lt	34	46	45	51
2nd Lt	30	45	46	57

It would take us too far afield to review this work here Perhaps the information in Table 60 1 (from 15), based on two samples totaling 7,771 cases, will suffice to indicate the type of evidence available to support the statement that it [a combination of forced choice and graphic rating scales embodied in an official efficiency report] produces ratings which are more valid indices of real worth' (10) This Table reports some of the results of a study conducted in connection with the regular reporting period of 30 June 1946 Both WD AGO Form 67 and WD AGO Form 67 1 were completed for the same officers The score on Form 67 was an average of ten 8 point graphic rating scales, the score on Form 67 1 was a combination of forced choice and rating scales The criterion used was an average of ratings by superiors, subordinates and associates ob

a distribution based on a scale of 220 used points (Form 67 1) with a distribution based on a scale of 43 used points (Form 67) [6] The attempt to equate the range for the purpose of comparing distributions on the two forms gives the traditional rating scale (form 67) every advantage

In the third place, Travers has missed the point A Personnel Research Section Report, dated 17 January 1947 (15) contains the information on which were based the illustrative Distributions reproduced as his Figure 59 1 The computation of the third and fourth moments contained in that report shows that Form 67 has greater leptokurtosis and that Form 67 1 has greater negative skew "This will mean that

[6] The score on Form 67 has a possible range of −4 to +7 Scores below 2 7 are rare Considering the score in tenths of a point gives 43 points in the actual range

Form 67 1 will be more discriminative of extreme cases than will 67, particularly at the low end of the distribution (15, p 7)

In the same Personnel Research Section Report, there appear data (15, p 16) which show the percentage of officers at two cut points on the distributions These data are reproduced as Table 60 2 If the equated data are taken at face value, this table clearly brings out the better discrimination by Form 67 1 at both ends of the distribution Below the point where the curves (23, Fig 1) cross at the low end of the distribution, there are 18 2 per cent for Form 67-1 and 9 5 per cent for Form 67 Above the point where the curves cross at the high end of the distribution, there are 10 7 per cent for Form 67 1 and 8 6 per cent for Form 67 Table 60 2 shows the same type of information by grade The point of particular interest is that for the lower grades, the difference in effectiveness of the two forms is most pronounced at the high end of the scale, for the upper grades, the difference is most marked at the lower end of the scale In the light of this kind of data, there is no question as to which form is the more useful

TABLE 60 2

Percentage of Officers Beyond Low and High Cut Points on Form 67 and Form 67-1

(From 15, Table V)

Grade	*% to Lower Cut from Bottom*			*% to Upper Cut from Top*		
	Form 67	*Form 67 1*	*% Excess*	*Form 67*	*Form 67 1*	*% Excess*
Colonels	31 46	60 28	28 82	No cross	No cross	0 00
Lt Colonels	24 69	48 15	23 46	No cross	No cross	0 00
Majors	13 28	33 65	20 37	3 32	4 53	1 21
Captains	12 06	20 24	8 18	5 07	10 10	5 03
1st Lts	3 56	12 93	9 37	10 32	17 22	6 90
2nd Lts	5 51	13 53	8 02	9 72	17 89	8 17
Combined	9 48	18 16	8 68	8 55	10 67	2 12

We disagree with Travers' conclusion "The data [from 14] suggest that experimentation with different types of directions may yield much more important results than experimentation with forced choice scales" (23, p 68) Experience of the Personnel Research Section indicates that this statement represents an extreme oversimplification of the problem involved In our experience, the basic attitudes of the raters, determined in large part by knowledge of the uses to which a rating scale is to be put, are little influenced by specific directions [7] If psychological effects of such a kind influence ratings, it would appear more likely that they will be brought about by the rater s actually doing something (for example, some form of preliminary ratings) than by a certain type of instruction Furthermore, raters do not necessarily follow the instructions Attention is called to the fact that despite what must be an extraordinary variety of directions for merit rating forms, high negative skew and leptokurtosis are almost invariably characteristic of the ultimate distributions These stubborn characteristics, in fact, have served as motivation for the search for other than the traditional rating techniques

We disagree with Travers' interpretation of a rater agreement represented by a correlation of 0 69 (23, p 69) While it is not the intention to discuss Travers comments on Richardson s study, from which this "reliability coefficient is cited it should be pointed out that the only 0 69 leads to the wrong evaluation of a coefficient of this magnitude This may be il

[7] The evidence for grade bias, i e, higher ranking officers being rated higher, in Table 60 2, despite careful instructions to disregard grade, is a case in point

lustrated from a follow up study of the validity of Form 67 1 (17) In this study, rater agreement on the criterion ratings obtained at a single sitting was represented by $r = 24$ Agreement between the official Form 67 1 raters for 914 cases was represented by $r = 56$ In Army experience at least, rater agreement as represented by coefficients as high as 70 is a rare finding and not to be considered unusually low [8]

Efficiency Reporting Problems Omitted From Travers' Article

In attempting a critical review of a technique, it is customary to discuss problems which it was hoped this technique might solve, and to cite the complete evidence These comments would seem to be particularly pertinent to articles appearing in the *Psychological Bulletin* We have already indicated deficiencies in the citation of evidence It will, perhaps, clarify the problem if a short history is given here

It is a truism in rating literature that halo, leniency, and rater differences in standards are basic problems The Army Officer Efficiency reporting system, as exemplified in WD AGO Form 67, had become increasingly subject to these influences [9] Figure 60 1 illustrates the increasing tendency for officers to be rated higher with the passage of time Form 67 was not believed by the Army to be serving its purpose, largely because it had lost its discriminating value at the high end of the scale The Army directed the Personnel Research Section to develop a rating system which would meet its needs to a greater degree The basis for the research on the problem of efficiency reporting was the study concerned with the development of procedures for the integration of officers into the Regular Army following

[8] Before leaving this section, a slight error should be corrected Travers attributes the quotation beginning 'a single over all rating on a 20 point scale ' on page 67, to his reference No 4 The quotation is actually contained in his reference No 5

[9] It is probably more accurate to say that Form 67 had become increasingly subject to halo and leniency No direct evidence is available concerning variations in rater agreement

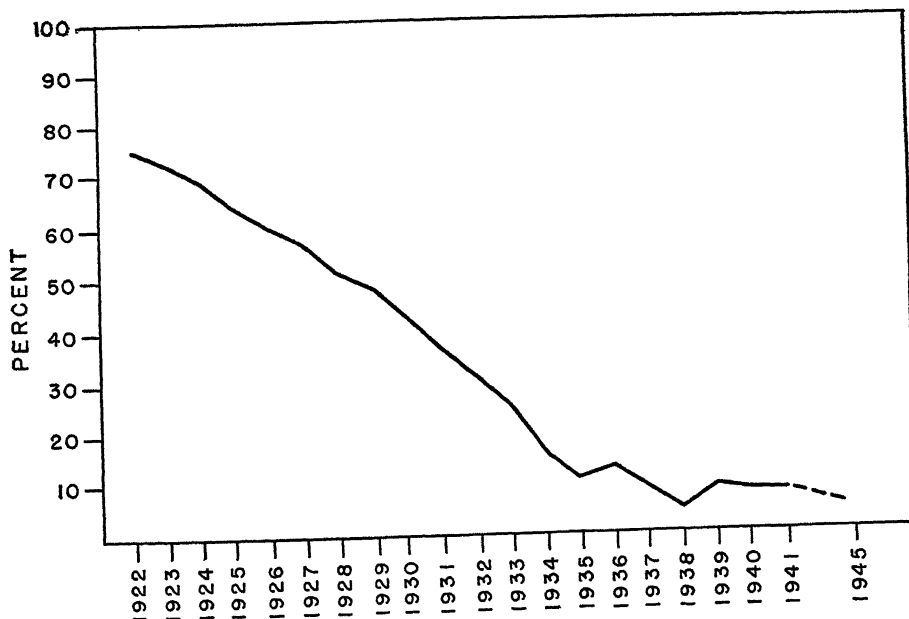



FIGURE 60 1 *Percentage of all regular Army captains receiving less than excellent ratings on WD AGO Form 67, 1922–1945 No ratings for war years 1942–1944*

World War II This research has been outlined by Richardson (5) The first Personnel Research Section studies of the forced choice technique were undertaken under this program[10]

Following this integration program, studies were initiated which compared five different rating forms, including rankings, various kinds of traditional type ratings, and forced choice The first series of studies involving some ten thousand officers in the United States and Europe showed a slight superiority for the forced choice rating form This superiority, coupled with the hypothesis that there would be less change in the distribution for the form involving the forced choice items when the form was used on an official basis, led to the decision to use it along with the old Form 67 at the June, 1946, rating period Details with respect to skew and means are presented elsewhere (22) Results bearing on the validity have already been presented in Table 60 1

This is not the place to review in detail the further findings of these and subsequent studies It seems more helpful at this point to summarize the advantages and disadvantages of the forced choice procedure as applied in the Army efficiency reports

The principal disadvantage is that the use of the technique had tended to be unacceptable to Army officers (although, apparently, more acceptable in industry) Acceptability is an especially important problem in rating procedures because of the effect on the rater s attitude Two comments may be made about the unacceptability

1 The name, forced choice, is an unfortunate one Reasons for its origin are readily understood in the light of the fact that the original presentation asked the individual to pick one of the two items as most descriptive of himself Actually, as previously suggested, forced choice might better be considered a scaled multiple choice item

2 The second point that should be made concerning the acceptability of the items arises out of the conversion of the raw scores on efficiency reports to standard scores Converting to a relative scale caused raters to feel that their ratings were not properly represented by a particular standard score, especially those below average In objecting to Form 67 1 there was much confusion between the effects of the forced choice technique and the effects of the use of a relative standard score scale

The advantages of the technique may be summarized as follows

1 It reduces halo Raters completing graphic ratings within the same form tend to mark them all pretty much the same way, i e, the correlation between graphic rating scales is high In completing two sections of forced choice items, raters likewise tend to mark them pretty much the same i e, forced choice sections also correlate high, but not as high as rating scales Forced choice ratings and rating scales correlate less than do rating scales with rating scales, or forced choice with forced choice ratings (17) On the simple basis of lower intercorrelation, a combination

[10] In the interest of historical accuracy, some elaboration of Travers statement concerning the origin of the forced choice technique should be made The idea of forced choice was suggested by Dr Paul Horst in a discussion at an APA meeting He himself does not remember the incident Dr Wherry was sufficiently interested to develop the scaling methods to achieve the process as Dr Horst had discussed it namely, to present, simultaneously items which looked alike to the individual completing a personality scale and yet had differing significances Dr Wherry developed the scaling procedures while working for the Civil Aeronautics Authority and brought them with him when he came to the Personnel Research Section Jurgensen (3) during this period had been working on a somewhat similar idea, a fact which was not known to the Personnel Research Section until after World War II It was Wherry s basic scaling technique which served as a point of departure for the work of the Personnel Research Section, first in the application to the development of personality inventories and later in the application to efficiency reporting As noted in this article the technique has been most successful in application to self rating, i e, personality inventories

of the two techniques would have greater possibility of increasing validity In comparison with Form 67, Form 67 1 has persistently yielded slightly greater validity, perhaps for this reason

2 It reduces bias, for example, it is less influenced by rank of the rater officer than was Form 67 (10) On the whole, the total score on Form 67 1 agrees better with an average criterion rating than do any of its sections (9, 17) The use of the average in itself is a conventional means of reducing bias

3 Forced choice item validities tend to be stable over a period of time From December, 1946, to January, 1949, item validities correlated from 50 to 60 (8) This stability is especially noteworthy in view of the narrow range of these item validity coefficients Although this evidence needs confirmation, it is the sort of evidence which encourages experimentation with the technique

4 Raters agree better on a report composed of both types of technique than they do on either type alone (17) This finding is of first importance Ratings on Form 67 1 for two successive rating periods for a group of 914 raters showed the best agreement for a combination of both techniques (i e, total score) than for either technique (Table 60 3)

mention in his discussion A further point which is illustrated in the above tabulation is that rating scales and forced choice ratings may differ among themselves It is therefore, difficult to make any hard and fast generalizations concerning either type of rating

The research of the Personnel Research Section, plus certain theoretical considerations, has persistently affirmed the value of the forced choice technique Since in the Army, at least, efficiency reports are usually considered for the entire career and since no technique or combination of techniques has brought rater agreement up to a satisfactory value our attention has been directed to developing a system whereby fluctuations in rating owing to leniency or other purely biasing factors might be reduced To put it another way, so much more is gained by combining ratings made by different raters than by improving the rating of a single rater through the use of a special technique that our emphasis is on averaging the reports prior to making use of them Obviously, for such a system to work, an adequate distribution of single ratings must be maintained From Figure 60 it is clear that over a period of time, ratings on the traditional type of rating scale in the Army tend to become restricted to the upper portion of the scale

TABLE 60 3

Agreement of Raters on Successive Reports, Form 67 1

Traditional rating sections	*r*	*Forced choice sections*	*r*		*r*
Section V	47	Section IV	42		
Section VII	39	Section VI	45	*Total Score*	56

In the absence of other than a rating criterion the problem of validity of efficiency reports may reduce to one of this kind of reliability Thus, the greater rater agreement on a combination of the two techniques is of special significance

It should be noted that the criterion used for the validity coefficients presented above was the average of a series of rating scales Rating scales would, therefore, be favored, a point which Travers neglects to

If this kind of trend can be established as characteristic it would appear to be necessary to develop techniques such as forced choice which have the promise of maintaining a spread in the ratings

It should be further noted that in the industrial situation where people may be rated consistently by the same rater, this averaging system probably will not work In such situations therefore, an effort to develop techniques such as forced choice is particularly needed

In conclusion, three points should be emphasized Travers discusses forced-choice technique as applied by itself It should be observed that both in this reply and in the research reports of the Section, the value of forced choice in combination with the traditional type of rating scales has been stressed Until an experiment is set up which gives each technique an equal chance to prove its worth—that is, the correlation not being subject to technique contamination—conclusions as to the value of techniques used will not be definitive

Secondly, the forced choice technique has been discussed in terms of the way it has been applied There are many ways in which it might be improved, e g, in method of calculating preference and dis criminative values, in method of pairing alternatives, perhaps by further application of suppressor theory or better grouping in terms of item content And, finally, the well demonstrated value of the technique in the construction of self rating scales of the personality inventory or self descriptive type is again stressed

BIBLIOGRAPHY

1 Brogden, H E, Newkirk, G F, & Loeffler, J C, The Prediction of Officer Potential of ROTC Cadets,' *American Psychologist* 1950, Vol 5, 360 (Abstract)

2 Campbell, J T & Rundquist, E A, 'Scale Items for Inclusion in Forced choice Rating Form, *American Psy chologist* 1950, Vol 5, 280 (Abstract)

3 Jurgensen, C E, 'Report on the 'Classification Inventory, a Personality Test for Industrial Use, *Journal of Applied Psychology* 1944, Vol 28, 445–460

4 Karcher, E K, & King S H, Effect of Number and Order of Ratings on Reliability and Validity,' *American Psychologist* 1950, Vol 5 333 (Ab stract)

5 Richardson, M W, Selection of Army Officers,' In *New Methods in Applied Psychology* Proceedings, Maryland Conference on Military Psy-chology University of Maryland, 1947, 79–85

6 Richardson, M W, 'An Empirical Study of the Forced-choice Perform ance Report, *American Psychologist* 1949, Vol 4, 278–279 (Abstract)

7 Rundquist, E A, 'Personality Tests and Prediction,' In D H Fryer and E R Henry, *Handbook of Applied Psychology* New York Rinehart and Co, 1950 Vol 1, 182–191

8 Rundquist, E A, Winer, B J, & Falk G H, 'Follow up Validation of Forced choice Items of the Army Officer Efficiency Report," *American Psychologist,* 1950, Vol 5 359 (Ab stract)

9 Schneider, Dorothy E, & Blackburn, J R, Validity of a Graphic Scale of Officer Efficiency *American Psy chologist* 1950, Vol 5, 359 (Ab stract)

10 Sisson E D 'Forced choice—the New Army Rating *Personnel Psychology* 1948, Vol 1, 365–381

11 Staff, Personnel Research Section, AGO The Forced choice Technique and Rating Scales, *American Psy chologist* 1946 Vol 1, 267 (Abstract)

12 Staff, Personnel Research Section, AGO *Construction and Selection of Items for the Biographical Information Blank* (*BIB*) PRS Report No 703, 7 July 1945

13 Staff, Personnel Research Section, AGO *Validation of Form E of the Bio graphical Information Blank* PRS Report No 716, 8 July 1946

14 Staff, Personnel Research Section, AGO *Comparison of the Rating Check List (RCL) and Forced choice List (FCL) Methods of Obtaining Ratings* PRS Report No 717, 9 July 1946

15 Staff, Personnel Research Section AGO *A Comparison of Officer Effi ciency Ratings Obtained With the WD AGO Form 67 and the WD AGO Form 67–1 at the Regular Reporting Period of 30 June 1946* PRS Report No 725, 17 Jan 1947

16 Staff, Personnel Research Section, AGO *Item Analysis and Development of Scoring Keys for the Leaders' Course BIB* PRS Report No 764, 1948

17 Staff, Personnel Research Section, AGO 1949 *Follow up Validation of WD AGO Form 67–1* PRS Report No 791 (In preparation)

18 Staff Personnel Research Section, AGO *Determination of Preference and Discrimination Values of Phrases for Preference Check List Sections of Alter nate Forms of Officer Efficiency Report* WD AGO Form 67 1 PRS Report No 846 (In preparation)

19 Staff, Personnel Research Section,

AGO *Development and Validation of the Self Description Blank for Predicting Leadership Qualities of ROTC Cadets* PRS Report No 859 (In preparation)

20 Staff, Personnel Research Section, AGO *A Rationale for Minimizing Distortion in Personality Questionnaire Keys* PRS Report No 868 (In preparation)

21 Taylor, E K, Carroll, J B, & Winer, B J, 'Validity of the Army's Officer Efficiency Report," *American Psychologist* 1949, Vol 4, 284 (Abstract)

22 Taylor, E K, & Wherry R W 'A Study of Leniency in Two Rating Systems," *Personnel Psychology* 1951, Vol 4, 39–47

23 Travers, R M W, 'Validity and Rationale of Forced choice Technique *Psychological Bulletin* 1951, Vol 48, 62–70

Critical Requirements A New Approach to Employee Evaluation *

JOHN C FLANAGAN

Based on a paper read at a joint meeting sponsored by the Psychometric Society and the Division of Evaluation and Measurement of the American Psychological Association at Denver, September 1949

Experience during the past few years has brought psychologists to a realization of the central role of measures of performance for personnel administration as well as personnel research Unless a satisfactory criterion measure is available research has been found to be not only worthless but in many instances definitely misleading Personnel actions without adequate personnel evaluations are little better than sheer guesswork

This situation has led to an intensive study of the fundamental nature of validity and criterion measures Inevitably this study has in turn forced investigators to a more careful examination of the definition of the specific activity on which research is being done Psychologists now see that without a definite and detailed definition of an activity or job in terms of actual behavior and the results of this behavior, the establishment of a criterion measure or personnel evaluation system is entirely out of the question Thus it becomes necessary to make an intensive analysis of the behavior of workers doing a job The usual techniques for job analysis were designed for a different purpose They could not be expected to carry this responsibility They provided hunches, opinions, and general descriptive materials In practice, the research worker found that taking such findings too seriously was likely to lead to serious error and job analysis results came to be regarded as a necessary preliminary step to be followed by systematic and thorough studies covering a very wide variety of materials In its extreme form this latter procedure was known as the "shot gun" approach With a wide enough scatter it was hoped that a few hits would be scored

A new approach has been developed which is designed to place a much heavier emphasis on the study of the behavior of the worker on the job It aims to collect representative samples of observed behavior which can be used as a basis for obtaining objective, quantitative data regarding the job It is hoped that instead of opinions and hunches, activity analysis can be made to yield the type of sampling data which can lead to inferences and predictions of testable reliability and validity

The essence of this new procedure is to establish the critical requirements of a job

* Reprinted from *Personnel Psychology*, Vol 2, No 4, Winter, 1949

or activity through direct observations by participants in or supervisors of the job or activity A critical requirement is defined as a requirement which is crucial in the sense that it has been responsible for outstandingly effective or definitely unsatisfactory performance of an important part of the job or activity in question Thus a critical requirement differs from the requirements which appear important but in practice have no important effect on performance with respect to the specified activity Observation of personnel engaged in a specific activity leads directly to critical requirements in terms of what workers actually do on the job In addition to such critical requirements in terms of behavior, it is desirable to determine critical requirements of the work in terms of aptitude, training, information, attitudes, habits, skills, and abilities Critical requirements of these latter types must be based on inferences and hypotheses These inferences and hypotheses may be checked by empirical studies

It is believed that the determination of critical requirements in terms of behavior is a necessary condition to an adequate definition of the job in terms of behavior Such a definition of the job must include the identification of the aspects of behavior to be included, standards of satisfactory performance for these, and estimates of their relative importance It is also clear that a definition of the job in terms of objectively described and evaluated behaviors of this type provides an almost complete statement of an adequate criterion measure of effectiveness on the job Similarly, no criterion or evaluation system which ignores these definitions of standards and of relative importance can be satisfactory Thus it follows that the problems of job definition, job requirements, and criteria of success necessarily reduce to one and the same problem, at least with respect to their major outlines

It should be emphasized at this point that observations of the behavior of the individual, or of the effectiveness of this behavior in accomplishing the desired results in a satisfactory manner, constitute not just one source of data but the only source of primary data regarding the critical requirements of the job in terms of behavior Neither outstanding ability nor unsatisfactory ability can exist independently of a series of observed behaviors Success and failure in the activity are nothing more nor less than a series of actions leading to observed results

Granting that objective data are greatly to be preferred to opinion, the next problem is how can satisfactory data be obtained Experience in establishing critical requirements indicates that five specific conditions must be satisfied These are as follows

1 It is essential that actual observations be made of the on the job activity and the product of such activity
2 The aims and objectives of the activity must be known to the observer Unless this condition is fulfilled it will be impossible for the observer or judge to identify success or failure For example, a foreman might be rated as very successful if the objective of his activity were taken as getting along well with the workmen under him At the same time he might be rated as very unsatisfactory if the objective is to produce materials
3 The basis for the specific judgments to be made by the observer must be clearly defined The data can be objective only if all observers are following the same rules All observers must have the same criteria for judging satisfactoriness The definition must clearly state whether or not a minor imperfection will be regarded as an evidence of failure or whether a product must be completely unusable to be classified as unsatisfactory
4 The observer must be qualified to make judgments regarding the activity observed Typically the supervisor on the job is in a much better position to make judgments as to whether or not behavior is outstanding or unsatisfactory, than is the job analyst or psychologist On the other hand the supervisor on the job is ordinarily quite lacking in the train

ing essential to make an inference as to the particular mental trait which caused the behavior to be successful or unsuccessful

5 The last necessary condition is that the situation be such that reporting is accurate The principal problems here are those of memory and communication It is also important that the observer s attention be directed to the essential aspects of the behavior being observed

In order that the critical requirements accurately reflect the data collected, the process of reducing several thousand specific observations of behavior to a fairly small number of critical requirements must be competently done The synthesis of the critical requirements from a variety of specific behaviors must be such that judges will agree that each of the specific behaviors should be classified under the summary statement which has been developed to include it For maximum usefulness the critical requirements also should be structured in such a way that they provide a coherent picture of the activity

In conclusion, a few of the devices which have been found effective in establishing critical requirements will be listed

1 The critical incident technique which consists in getting incidents of extreme behavior, either outstanding or unsatisfactory, has been found very effective in collecting data where adequate records regarding behavior data are not available This procedure has considerable efficiency because of the use of only the extremes of behavior It is well known that extremes can be more accurately identified than behavior which is more nearly average in character It must be verified that the five conditions noted above are satisfied so that there will be no biasing of the sample of incidents, due either to selective memory or to inadequate definition of the type of incident to be included

2 A second device which has been found very helpful is the evaluation and classification of incidents at the time of observation It is much simpler to evaluate a sample of behavior at the time it is observed, when all relevant details may be noted or checked, than it is at some later time when further examination of the behavior is impossible If an incident is evaluated and classified at the time of its occurrence and a mental note made that it is to be recorded later, the recall and recording of this material is greatly improved in accuracy and the time necessary is considerably reduced

3 Another device which has been found to save much time is the preparation of a complete observational record form which contains practically all of the types of incidents which are likely to be observed Incidents may be reported on such a form by merely tallying in the appropriate space

4 A final use of such data is the direct conversion of frequencies into statistical estimates for purposes of prediction and evaluation In situations where the observations can be accurately classified and where also an adequate representative sample of behavior can be obtained it is possible to obtain fairly accurate, unbiased estimates of the importance of a particular critical requirement with relation to the other requirements Where the requirements are independent this can be converted into a correlation coefficient by using the usual formula for estimating the correlation attributable to common elements in the variables

The critical incident technique has been successfully used in a number of situations One of the first uses was in establishing the critical requirements for the United States Air Force officer In this instance the incidents were obtained by interviewing officers who had had a considerable amount of military experience The procedure has been used in determining the critical requirements for research workers in scientific laboratories In this situation, also, experienced supervisory personnel were requested to provide the incidents Recently the procedure has been applied to hourly wage workers in one of the divisions of a large industrial corporation In this situation the foremen supplied the incidents from which the critical requirements were developed The Committee on Ethical Standards of the American Psychological Association is using a variation of the critical incident technique in ob-

taining a survey of practical problems in that area

In the course of these studies a substantial amount of research and developmental work has been accomplished and it is believed that there has now been sufficient experience with the critical incident technique to demonstrate that the critical requirements of a job can be established through direct observations of personnel engaged in the activity

The adoption of these techniques integrates the problems of job definition, selection and classification, and the development of criterion measures and makes it possible to carry out research on the criterion problem on a sound and rational basis

REFERENCES

Flanagan, John C, A New Approach to Evaluating Personnel' *Personnel* 1949, Vol 26, 35-42

Flanagan, John C, *Critical Requirements for Research Personnel A Study of Observed Behaviors of Personnel in Research Laboratories* Pittsburgh American Institute for Research, 1949

Flanagan John C, Job Requirements, *Current Trends in Industrial Psychology* Pittsburgh University of Pittsburgh Press, 1949

Flanagan, John C, *AAF Aviation Psychology Program Research Report No I* Washington U S Government Printing Office, 1948

Gordon, Thomas, *The Airline Pilot A Survey of the Critical Requirements of His Job and of Pilot Evaluation and Selection Procedures* Washington Civil Aeronautics Administration, Division of Research, Report No 73 1947

Nagay John A, *The Development of a Procedure for Evaluating the Proficiency of Air Route Traffic Controllers* Washington Civil Aeronautics Administration, Division of Research, Report No 83, 1949

Preston Harley O, *The Development of a Procedure for Evaluating Officers in the United States Air Force* Pittsburgh American Institute for Research, 1948

The Development of a Procedure for Evaluating Officers in the United States Air Force *

The research described below led to the recent adoption, by the United States Air Force, of a new procedure for evaluating its officers The study was initiated because of the need within the military services for accurate information about the relative effectiveness of officers Although the research dealt with the effectiveness of officers the Critical Incident Technique used and the Observational Record of Work Performance developed will be of wider interest because of their applicability to other types of personnel and personnel research

The aim was to develop a practical, simple direct evaluation form to provide early identification of officers both for elimination and for accelerated advancement The objective of the research was to develop a practical way to identify those commissioned officers of probationary status within the Air Force whose performance indicated either very high or very low promise of future success as officers Impetus for the project and several general specifications for the final evaluation procedure came from the Officer Selection Committee which had the responsibility for making final recommendations for the integration of temporary officers into the regular Air Force Previous rating procedures had been found to be unreliable The committee offered three recommendations for an acceptable evaluation procedure (1) it should be designed for use within the existing framework of the Air Force, (2) it should be simple and direct

* Reprinted from *Research Notes*, No 2, June 1949 Pittsburgh American Institute for Research

to administer, and (3) it should yield a numerical score

The first and most basic problem was the determination of a well defined and delimited standard for judging officers, without which no evaluation procedure would be of value The second problem was how best to measure the individual officer against such a standard

The traditional "ideal officer approach was discarded in favor of a direct analysis of job requirements in terms of behavior The traditional approach to officer evaluation has been through the agreement of responsible authorities on personal traits which ideally every officer should possess Thus, by defining an 'ideal" officer, the standard for all other officers was presumably set However, it was found that this general approach has certain limitations, in this project, many of those limitations were avoided through analysis of the job of an Air Force officer by a method designated as the *Method of Critical Requirements*

In this method, the basic data consist of behavioral descriptions of performance on a specific job Such descriptions were obtained in this study by interviewing officers in a position to judge the effective or ineffective performance of other officers This method lays stress on those job requirements which are *critical* in the sense that they have been deciding factors in judging a significant number of individuals as "successful" or "unsuccessful' on a specific job Emphasis is thus shifted from qualities of 'goodness" to specific ways of acting which are observed to be effective, and the "good officer is defined as one who effectively meets the important demands of his job—that is, one who meets the critical requirements of an Air Force officer

While several techniques are possible with the method of Critical Requirements, the most useful one was found to be the "Critical Incident Technique" In this technique the attention of the officers interviewed was focused first on a specific situation in which they had judged the effectiveness of another officer, and then on the specific behavior of that officer By the use of carefully phrased questions, detailed descriptions were obtained of how an officer acted in a particular job situation which caused him to be judged as effective or ineffective in that situation Both the phrasing and sequence of the questions were field tested prior to actual use

Over 600 officers provided 3 000 descriptions of outstanding and unsatisfactory job performances A total of 640 officers were interviewed, either individually or in small groups of from three to seven, at 16 locations in the United States At any one location, officers from all the different organizations were included The rank of officers interviewed ranged from Lieutenant through General The interviews were concentrated among field grade officers, since early experiments showed that junior officers could relate fewer incidents than senior officers of ineffective behavior which they had directly observed The interviewing was continued until a study of the incidents being obtained revealed that additional interviews would result largely in repetition of information already received A total of 3,029 incidents were obtained, 1,228 covering effective behavior and 1,801 covering ineffective behavior

Many incidents covered more than one specific behavior An analysis of the incidents resulted in 2,142 effective behaviors and 2,869 ineffective behaviors As each incident described an officer performing his duty in such a way that other officers judged him to be especially effective or ineffective, each was potentially a critical requirement of an Air Force officer However, some of the reported behaviors were identical, and some were similar in essential respects As a first step, all identical behaviors were grouped In subsequent steps, groupings were made of those behaviors which had occurred in related situations Each group of similar officer behaviors could then be considered as descriptive of a critical requirement The purpose of the project—to develop officer evaluation procedures—made this type of analytical reduction necessary, since it would be impractical to evaluate officers on several hundred possible behaviors Continual revision was made until six major areas and 58 sub areas or critical

requirements were evolved These sub areas covered the basic requirements, yet fell within the range of practicability

Figure 62 1 gives the percent of total out standing and unsatisfactory behaviors in the pre tryout phase Subsequent changes were minor Area headings conform to those in the form adopted for use by the Air Force

In order to select a final form and a method for using the 58 critical require ments in an evaluation procedure, 12 dif ferent types of experimental procedures were given brief field tryouts From these, and from conferences with civilian and military specialists, an outline emerged of the type of procedure which merited fuller development and extensive tryout in the Air Force In general, it consisted of matching the facts of observed perform- ance over a period of several months with a series of descriptions of how effective and ineffective officers had met or had failed to meet the critical requirements of an Air Force officer Two standard forms were required a booklet containing these descriptions, and a report form for the de- grees of effectiveness assigned at the end of a reporting period Specifications set up for the workbook to be used by evaluating officers called for presenting each critical requirement as a five point scale, with a description of how an officer acts when succeeding or failing to meet that require ment arranging the scales so that they would all run from ineffective behavior at the low end to very effective behavior at the high end, describing at the central point of each scale an acceptable way of acting, and grouping the scales under areas for which separate scores could be ob tained

Nearly 2 000 officers took part in the tryout of he tentative evaluation forms The preliminary forms of the evaluation procedure were given a field tryout on nearly 2,000 officers, divided into two groups of approximately equal size Quotas were assigned to each major command A

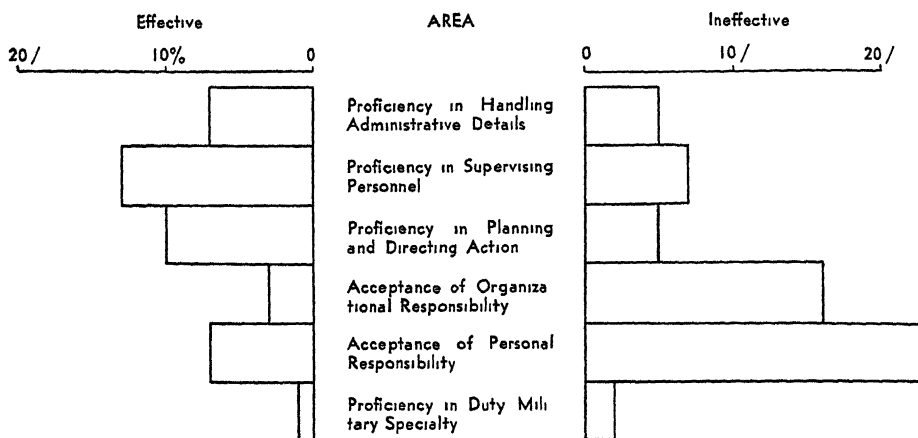



FIGURE 62 1 *Per cent of total effective and ineffective behaviors falling within each area*

project officer in each major command then allocated its quota to various sub commands and stations The project officer (the personnel officer in most commands) selected the officer in charge of the tryout at each station Each reporting officer was required first to select the two probation ary officers holding the highest temporary rank or grade among all such officers under his supervision Then, the officer with the lower serial number was evaluated in the first phase of the tryout, and the one with the higher number in the second phase In the first phase, the reporting officer made an immediate report on the effectiveness of the officer being evaluated In the sec ond phase, a report was made on the offi cer after he had been observed closely for a month Thus, the procedure was tested

as a single record type of report, and also as a workbook for the recording of obser vations

The officers using the form in the tryout favored its adoption for general Air Force use Reporting officers were asked to com ment on the procedures Of the 767 com ment sheets received 709 were classified according to the reactions of these officers to the procedure Five hundred sixty nine felt that the procedure should be adopted in its tryout form, or after minor revisions had been made, and 140 had several crit icisms of the form or of this general ap proach to evaluation Another interesting comparison was between the new pro cedure and previous systems Three hun dred sixteen officers volunteered such comparisons, 92 per cent stated the new procedure was better, 3 per cent that it was no better, and 5 per cent stated it was worse than the old systems

By assigning numerical values to the de scriptions along the scale, it was possible to arrive at a numerical score for each of the areas and for all the areas combined All such scores were converted to stanine scores Partial reports could be made, when necessary, and further flexibility was achieved by separating the formal report of officers effectiveness from the check list on which observed performances were matched with descriptions of effective be haviors

The tryout led to certain minor revisions in the evaluation form The tryout form contained 58 items, the form adopted for use contains 54 Certain items were com bined because they appeared to be very similar under the actual conditions of use

The evaluation form containing the 54 critical requirements of an effective Air Force officer, is now being used by the USAF Major areas and sub areas of the evaluation procedures as adopted by the United States Air Force are indicated be low

I *Proficiency in Handling Administrative Details*
1 Understanding instructions
2 Scheduling work
3 Getting information from records
4 Getting ideas from others
5 Checking accuracy of work
6 Writing letters and reports
7 Getting cooperation
8 Presenting finished work
9 Keeping records
10 Keeping others informed
11 Rendering effectiveness reports

II *Proficiency in Supervising Personnel*
12 Matching personnel and jobs
13 Delegating authority
14 Giving orders and instructions
15 Insuring comprehension
16 Giving reasons and explanations
17 Supporting authorized actions
18 Encouraging ideas
19 Developing teamwork
20 Setting a good example
21 Assisting subordinates in their work
22 Evaluating subordinates work
23 Looking out for subordinates' welfare
24 Maintaining relations with subordinates

III *Proficiency in Planning and Directing Action*
25 Taking responsibility
26 Solving problems
27 Making use of experience
28 Long range planning
29 Taking prompt action
30 Suspending judgment
31 Making correct decisions
32 Making forceful efforts
33 Absorbing materials

IV *Acceptance of Organizational Responsibility*
34 Complying with orders and directives
35 Accepting organizational procedure
36 Subordinating personal interests
37 Cooperating with associates
38 Showing loyalty
39 Taking responsibility for subordinates

V *Acceptance of Personal Responsibility*
40 Attending to duty
41 Attending to details
42 Reporting for appointments
43 Meeting commitments

44 Beıng faır and scrupulous
45 Maıntaınıng mılıtary appearance
46 Adaptıng to assocıates
47 Adaptıng to the job
48 Conformıng to cıvıl standards

VI *Profıcıency ın Duty Mılıtary Occupatıonal Specıalty*
49 Possessıng fundamental traınıng
50 Improvıng effectıveness
51 Keepıng well ınformed
52 Applyıng traınıng and ınformatıon
53 Showıng ıngenuıty ın specıalty
54 Handlıng related assıgnments

A full report ıs contaıned ın Harley O Preston *The Development of a Procedure for Evaluatıng Officers ın the Unıted States Aır Force* Amerıcan Instıtute for Re search, Pıttsburgh 13, Pa, 7 July 1948, and ın Technıcal Appendıces and Notes ıssued separately

Thıs study was sponsored by the Officer Selectıon Commıttee and the Avıatıon Psychology Program of the Surgeon Gen eral of the Unıted States Aır Force

The Development of a Method of Evaluatıng Flyıng Skıll *

THOMAS GORDON

Thıs study was carrıed out when the wrıter was Dırector of Avıatıon Research of the Amerıcan Instıtute for Research It was made possıble by a grant from the Cıvıl Aeronautıcs Admınıstratıon under auspıces of the Commıttee on Avıatıon Psychology of the Natıonal Research Councıl The wrıter wıshes to express apprecıatıon to John C Flanagan, Presı dent of the Amerıcan Instıtute for Research, for hıs ınvaluable guıdance, assıstance and en couragement durıng all stages of the study The wrıter ıs grateful to hıs staff of able research workers and clerıcal assıstants The full report has been publıshed as CAA Dıvısıon of Research Report No 85, Aprıl 1949

INTRODUCTION

The need for accurate measurement of flyıng skıll exısted durıng the last war and has sınce become ıncreasıngly crıtıcal wıth the rapıd ıncrease ın cıvılıan flyıng, both prıvate and commercıal The serıous need for, as well as the dıfficultıes ınvolved ın, developıng crıterıon measures of pılot pro fıcıency, however, repeatedly have been affirmed by many ınvestıgators ın thıs field (23, 6, 18, 11)

Prevıous research aımed at the develop ment of more objectıve methods of eval uatıng flyıng skıll has been concentrated ın the mılıtary servıces and ın prıvate flyıng The problem of gettıng ımproved pılot evaluatıon procedures accepted by such agencıes, however, has troubled most of the prevıous ınvestıgators ın thıs field Although several studıes have reported the development of ımproved pılot profıcıency measures and although psychologısts have known for at least ten years that tradı tıonal methods are unrelıable and non dıscrımınatıve, all of the major agencıes employıng pılot evaluatıon procedures are stıll usıng the same basıc method—sub jectıve ratıngs Studıes of the subjectıve type of method consıstently have demonstrated that ıt does not result ın a satısfac tory amount of agreement between check pılots ındependently evaluatıng the profıcıency of the same group of pılots, ıt does not satısfactorıly dıscrımınate between relatıve profıcıency ın dıfferent aspects of flyıng, ıt does not gıve adequate ranges of the abılıtıes of dıfferent pılots, and ıt does not adequately predıct success ın later stages of traınıng (2, 4, 12, 18, 21, 9, 10, 3) Such research has led to repeated at tempts to develop ımproved pılot evalua tıon procedures of three general types (1) graphıc and photographıc methods,

* Reprınted from *Personnel Psychology* Vol 3, No 1, Sprıng, 1950

(2) rating methods, and (3) objective observation methods

From the studies in which graphic or photographic methods were developed or utilized (24, 19, 1, 32, 15, 16, 17, 29, 26, 27, 28), the following conclusions seem warranted

1 The methods of recording performance are highly objective, thus useful for research purposes

2 The methods require special equipment which is excessively costly

3 The records of the flight are not immediately available because of the time needed to print the film or analyze the records

4 The measures are not comprehensive, i e, do not cover all of the critical aspects of flying

5 The records themselves do not yield a measure of proficiency, hence, it is necessary to employ methods of evaluating the records The observer observer reliability of these evaluation methods has not yet been established for graphic records and has been established for photographic records on an extremely small sample of raters

6 Test retest reliabilities have not been adequately determined for either graphic or photographic methods in a situation where ratings of the two flights are made by different raters and without knowledge of the pilots' performance on the first flight

7 None of the studies has established the relevance of the methods with respect to flying proficiency in general

The findings from the studies in which rating methods were developed or employed (12, 25, 13, 30, 5, 20, 22) may be summarized as follows

1 No rating scale has been shown to have adequate test retest reliability

2 The relevance of rating scale types of procedures has not been adequately established nor have such methods been developed from careful job analyses

3 Only one rating type method has been developed which seems comprehensive, but studies have not demonstrated its reliability or relevance

The studies in which objective observation methods have been developed or evaluated (4, 18, 19, 28) indicate that this is the most promising type of procedure The method requires that numerous objective observations be made by check pilots during a standard flight, that the check pilot record his observations on standard forms immediately after the observation is made, and that scores be assigned to small segments of the pilot s performance rather than to the performance as a whole The research on such objective methods, however, has shown that reliable test *items* can be developed but more research is needed to develop a *single comprehensive flight check* combining many objective items Furthermore, these studies indicate that improvements in the procedures are needed in order to make them more acceptable to those whose responsibility it is to use the procedures

The researches sponsored by the Committee on Aviation Psychology (4, 19, 28) and those carried out by the AAF Aviation Psychology Program (18) have already made progress in the development of such procedures, despite the difficulties inherent in the problem These difficulties are related to the problem of satisfying the criteria of *reliability, relevance* and *acceptability*

1 Reliability Procedures with adequate reliability have been difficult to devise for a number of reasons, chief of which are difficulties of communication between pilot and check pilot in modern aircraft, atmospheric variables affecting plane performance from one flight to the next, difficulties of recording performance during flight, differences in the standards and judgment of check pilots

2 Relevance Getting relevant tasks into an evaluation procedure has been difficult largely because of the following the difficulty of simulating in the test situation all of the actual conditions encountered in flying, the prohibitive cost of long evaluation flights in high powered aircraft, disagreement among "experts" as to which tasks are the most relevant

3 Acceptability Achieving the acceptance of objective procedures has been dif-

ficult for many reasons, such as the fact that researchers have not always dealt with resistance to change as skillfully as they have dealt with the measurement problem itself, most of the procedures have been too difficult to administer in flight, investigators have not developed objective procedures which can be used with a number of different types of aircraft, measures of critical skills have often been left out of flight checks because of the difficulty in measuring them 'objectively", pilots have objected to complicated scoring procedures usually incorporated into objective flight checks, in striving for objectivity investigators have frequently constructed items requiring the check pilot to use points of reference which are not used by him in actual practise

The objective of the present study was to develop a single comprehensive flight-check which would be more reliable than currently used subjective procedures yet would measure the skills which are most relevant to success on the job and would be acceptable to those who eventually would use it It was decided to develop the flight check for the Airline Transport Rating flight examination, the one required of pilots in order to become certified as airline pilots

Principal Methods and Results

Determining the critical requirements of the job The problem of developing a flight check which would measure the most relevant skills was approached through employing the critical incident technique for determining the critical requirements of the airline pilot's job This approach is essentially one of utilizing reports of observed behaviors which have been shown to be critical from the standpoint of successful or unsuccessful performance on the job The method was used in the AAF Aviation Psychology Program in job analysis studies (31, 14) and has been described by Flanagan (6) This particular job analysis approach yields data in the form of observed behavior rather than lists of activities and traits based on opinions and judgments as typically obtained from other job analysis methods In this study there were three principal sources of data about the critical requirements of the job which were used as the basis for the development of the evaluation procedure These are summarized in Table 63 1

Accident reports were copied from the actual files of the CAB The pilot incidents were obtained through interviews with airline pilots in 18 different cities and from 27 different airline companies Interviewers used questions devised specially to elicit actual accounts of situations encountered by the pilots being interviewed—situations which had been brought on by some kind of ineffective behavior on their part or situations in which they felt they behaved ineffectively Flight-check incidents were those collected from check pilots in which they reported critical incidents in which pilots taking regular flight examinations behaved in such a way as to warrant a

TABLE 63 1

Sources of Critical Incidents

Critical Incidents	*Source*	*Number of Interviews*	*Number of Incidents Collected*	*Number of Incidents Used*
Airline accidents	Civil Aeronautics Board Accident Files	—	185	121
Pilot incidents (Near-accidents)	Interviews with Airline pilots	270	601	395
Flight check incidents (Near accidents and reasons for failure)	Interviews with Airline and CAA check pilots	58	137	137
Totals		328	923	653

TABLE 63 2

Critical Components of the Job of Airline Pilot as Determined From Ineffective Acts Extracted from Accident Reports, Pilot Incidents and Flight Check Incidents

Critical Job Components	*Frequency of Ineffective Acts in*			
	Accidents	Pilot incidents	Flight check incidents	Total
Planning and Preparing for Flight				
1 Obtaining information about conditions to be encountered in flight	0	7	3	10
2 Checking on the condition of the airplane and its equipment prior to flight	1	7	1	9
Controlling the Flight of the Airplane Within Prescribed Limits in the Performance of Routinely Used Maneuvers				
3 Taxiing	3	1	2	6
4 Taking off under normal conditions	6	13	5	24
5 Taking off under conditions of reduced visibility	0	3	3	6
6 Taking off under cross wind conditions	0	2	0	2
7 Making a contact approach and landing under normal conditions	16	85	33	134
8 Making a contact approach and landing under conditions of reduced visibility	25	40	20	85
9 Making a contact approach and landing under cross wind conditions	14	54	5	73
10 Making instrument approaches by means of reference to different types of radio aids	0	10	23	33
11 Recovering from a missed instrument approach or missed landing	1	3	3	7
12 Other maneuvers	0	4	1	5
Controlling the Flight of the Airplane Within Prescribed Limits Under Unusual Emergency Conditions				
13 Recovering from sudden engine failure and performing maneuvers with an engine out	3	3	23	29
14 Operating the airplane when the air is turbulent when runways are slippery, when icing conditions are present, etc	23	11	2	36
15 Controlling the airplane in unusual attitudes or at minimum airspeeds	0	3	20	23
Employing Procedures to Locate or Keep Track of Position in Flight or to Fly a Prescribed Course				
16 Navigating and orienting	7	13	41	61
17 Communicating with traffic control personnel	0	7	2	9
Operating Equipment of Plane and Carrying out Cockpit Procedures				
18 Remembering to carry out certain prescribed or appropriate tasks in connection with the operation of the equipment of the airplane	7	19	11	37
19 Operating the controls, dials and switches of the plane's equipment in a correct manner	8	17	28	53
Adhering to Prescribed Policies or Regulations and Taking Precautions Consistent with Safety				

TABLE 63 2—*Continued*

Critical Job Components	*Frequency of Ineffective Acts in*			
	Acci dents	Pilot inci dents	Flight check inci dents	Total
20 Conforming to regulations and policies	5	21	4	30
21 Keeping a constant lookout for possible collision objects and remaining attentive and alert	13	8	3	24
22 Taking special precautions or remaining on safe side	2	11	4	17
Remaining Emotionally Organized and Working Efficiently with Others				
23 Remaining emotionally organized in emergency situations	0	8	7	15
24 Working efficiently with other crew members	0	4	1	5
Total	134	354	245	733

TABLE 63 3

Frequency of Ineffective Behavior Occurring in Critical Incidents Involving Approaches and Landings Under Low Visibility Conditions

Ineffective Behavior	*Frequency of Ineffective Acts Obtained from*			
	Accidents	Pilot incidents	Flight check incidents	Total
Failing to align with runway or flying in correct heading from station to field	0	11	9	20
Failing to keep within sight of field	0	2	0	2
Failing to locate field after becoming contact, mistaking landmark for field	0	3	0	3
Failing to hold constant altitude when circling field	2	2	2	6
Failing to hold proper glide angle in descent	17	10	6	33
Failing to hold proper airspeed in descent	0	4	2	6
Failing to go around after overshooting	1	3	0	4
Turning too steeply when close to ground	0	1	1	2
Leveling off too high or too low	1	2	0	3
Flying partially instruments and partially contact instead of one or the other	0	1	0	1
Failing to stay aligned with runway on roll	1	0	0	1
Landing in field adjacent to airport	1	0	0	1
Landing downwind	2	0	0	2
Failing to plan approach	0	1	0	1
Totals	25	40	20	85

failing grade or to require the check pilot to take over the controls because of a crit ical situation caused by the examinee [1]

The 653 critical incidents were subjected to a content analysis which involved first extracting from each incident the specific ineffective pilot behaviors, then sorting these into separate job components, such as landings, takeoffs and taxiing, depending on where the incident occurred Finally, the ineffective pilot acts in each job com ponent were sorted into categories of sim ilar acts Table 63 2 presents the 24 differ ent job components into which the inef fective acts were classified and the fre quency of these acts As an illustration of the kinds of ineffective acts classified under the job components, the categories of acts under "Making a Contact Approach and Landing under Low Visibility Conditions" are presented in Table 63 3

The Development of the Evaluation Procedure The job of constructing the evaluation procedure involved the follow ing steps (1) selecting job related tasks identical or similar to the 24 job compo nents found most critical in the job an alysis, (2) arranging these efficiently into a standard flight in such a way that the flight check could be administered in the shortest possible time, (3) breaking down each job related task into the critical ob servable behavioral units derived from the job analysis and devising items for each behavioral unit, (4) devising the flight check form on which the check pilot records his observations

In order to make the flight-check form usable with different types of airplanes, a special kind of item was developed which allows the check pilot to write in different limits of air speed, altitude and heading which are appropriate to that particular type of airplane being used for the ex amination (see Item 5, Figure 63 1) Other features were incorporated into the flight-check form in order to solve the problem of having an evaluation procedure which is comprehensive yet can be used in the air without diverting too much of the check pilot s attention from keeping a close watch for other traffic or taking too much of his time to record observations A spe cial format was used for the flight check form which separates different items and clearly labels each The most distinctive feature, however, was the frequent use of graphic and pictorial items (see Items 2, 3, 4 and 6, Figure 63 1) Not only did these items decrease the amount of record ing time but it is felt that they contributed greatly toward making the flight check more reliable largely because they pro vided a much more objective definition of the limits of performance on an item than is provided by words alone

The final form of the flight check con sisted of 18 different tasks (or maneuvers), each of which contained from 2 to 9 items A sample maneuver is reproduced in Figure 63 1

The experimental try out of the evalu-ation procedure The first try out of the flight check involved administration to 27 Air Force pilots on two successive flights on different days During each of the two flights two check pilots made independent observations, the two check pilots on the second flight being different from those on the first Thirty different check pilots were involved, all of whom were experienced instrument instructors Approximately 15 different airplanes, all TB 25 s, were em-ployed for the try out Each of the 27 pilots was tested by four different check-pilots, referred to as observers A, B, C and D Observers A and C were the right seat observers on the first and second day re-spectively and observers B and D were the jump seat (the seat behind the right seat) observers on the first and second day re spectively The obtained observer observer reliabilities (AB and CD) and the test retest (ride ride) reliabilities (AC, AD, BC, BD) are presented in Table 63 4, top half

The second try out involved testing 26 CAA pilots using the same procedure as in the first try out A revised flight check form was used for this try out Revisions were made on the basis of (1) an item

---

[1] This phase of the study is described in more detail in T Gordon, The Airline Pilot s Job, *Journal of Applied Psychology* 1949, Vol 33, 2, 122–131

INSTRUMENT TAKEOFF

TELL EXAMINEE Line the plane up with the runway yourself When you have it the way you want it hold it with brake until I give the signal for takeoff

(1) POWER APPLICATION
- ☐ Smooth and positive
- ☐ Jerky or hesitant
- ☐ Excessively rapid

(2) HEADING ON ROLL
- ☐ Straight path only slight deviations stayed in center
- ☐ Large deviations yet stayed in center of runway
- ☐ Off toward edge of runway yet fairly straight path
- ☐ Off toward edge of runway plus large deviations

(3) ATTITUDE AT END OF ROLL
- ☐ Normal
- ☐ Too tail low
- ☐ Too tail high held on ground too long

(4) FLIGHT PATH JUST AFTER BECOMING AIRBORNE
- ☐ Pulled up steep
- ☐ Normal
- ☐ Held down
- ☐ Dropped back

(5) AIRSPEED IN CLIMB
Slow | Within Limits | Fast
(____) -10 mph (____) Recommended Climb Speed (____) +10 mph

(6) HEADING IN CLIMB
- ☐ Fairly straight path
- ☐ Off to either side

(7) CHECK PILOT ASSISTANCE
- ☐ Assistance not necessary
- ☐ Assistance necessary

COMMENTS

FIGURE 63 1 *Flight-check form*

analysis of the first form, (2) maneuver reliabilities determined from first try-out, and (3) suggestions of the check pilots in the first try out Revisions consisted of adding more pictorial items, changing the sequence of several maneuvers, improving the directions for check pilots, eliminating several items objectionable to the check-pilots In this try out CAA examiners responsible for administering the Airline Transport flight exam acted as both check pilots and examinees Tests were conducted in the DC 3 and DC 4 airplanes Obtained reliabilities are shown in the bottom half of Table 63 4 It can be seen that the high observer observer reliability obtained in the first tryout held up on the second and that the test retest reliability improved, probably as a result of the revisions of the form

Empirical proof of the extent to which the critical incident approach produced a relevant flight check will be obtained in a third try out planned with already qualified airline pilots, on whom there are more adequate criterion data against which to validate the flight check In this study there were no adequate measures of proficiency available on either the AAF or the CAA pilots

An analysis was made of the CAA check pilots' written responses to a questionnaire in order to obtain an indication of the degree to which the flight-check satisfied the criterion of acceptability In general, although they saw both advan

TABLE 63 4

Observer Observer and Test Retest Reliabilities of the Flight Check on First Try out

	Number of Pilots	Reliability Coefficients							
		AB	AC	AD	BC	BD	CD	O_1O_2*	R_1R_2*
First try out Total Flight check score	27	87	49	49	59	50	76	82	52
Second try out Total Flight check score	26	84	81	75	74	74	87	86	76

* O_1O_2 represents the combined observer observer coefficient (AB and CD) and R_1R_2 represents the combined test retest coefficient (AC, AD, BC and BD) These combined coefficients were calculated by means of the z transformation technique

tages and disadvantages, the majority were either in favor of adopting it as the regular authorized flight exam or in favor of continuing work on it to make it ready for adoption

Conclusions

1 The flight check developed in this study can be considered a reliable procedure for arriving at an overall evaluation of pilots with experience similar to those tested in this study Both observer observer and test retest reliabilities were considerably higher than those reported for other methods of evaluating pilots

2 The flight check can be considered relevant to success or failure on the job to the extent that it is made up of tasks which are identical or similar to the requirements of the job found to be critical by job analysis More acceptable evidence of its relevance must await further try out

3 The majority of check pilots of the kind who will use the procedure consider it an acceptable method, although further revisions were recommended

4 Several distinctive features of this flight-check differentiate it from previous objective flight checks and probably contributed greatly to its high reliability and/or its acceptability

a A new type of graphic or pictorial item was used frequently

b A new type of item was devised which makes it possible to use the same flight check form for different types of planes, whereas previous objective flight checks were limited to use with only a single type of plane

c The items adequately cover the most critical aspects of piloting, because the flight check was built upon the findings of the critical requirement study

References

1 Backstrom, O, Jr, and Viteles, M S, *An Analysis of Graphic Records of Pilot Performance Obtained by means of the R S Ride Recorder* Washington CAA Division of Research, Report No 55, 1946

2 Cowles, J T, and Dailey, J T, 'The Measurement and Prediction of Civilian Flying Instructor Proficiency," *American Psychologist* 1946, Vol 1, 292

3 Crawford, M P, and Dailey J T, An Analysis of Elementary Pilot Performance from Instructors Comments, *American Psychologist,* 1946, Vol 1, 292

4 Edgerton, H A, and Walker, R Y, *History and Development of the Ohio State Flight Inventory Part I Early Versions and Basic Research* Washington CAA Division of Research, Report No 47, 1945

5 Festinger, L, Kogan, L S, Odbert, H S, and Wapner, S, *An Analysis of Inspectors' Ratings of Check Flights as Recorded on Form ACA 342Z* Washington CAA Division of Research, Report No 58, 1946

6 Flanagan, J C (ed), *The Aviation Psychology Program in the Army Air Forces* Army Air Forces Aviation Psy

chology Program Research Report No 1, Washington U S Government Printing Office, 1948

7 Gordon, T, *The Airline Pilot A Survey of the Critical Requirements of His Job and of Pilot Evaluation and Selection Procedures* Washington CAA Division of Research, Report No 73, 1947

8 Guilford, J P, *Fundamental Statistics in Psychology and Education* New York McGraw Hill Book Company, Inc, 1942

9 Henneman, R H, "Proficiency Measures for Fighter Pilots at the Operational Level of Training in the Army Air Forces," *American Psychologist*, 1946, Vol 1, 293

10 Henneman, R H, Hausman, H J, Mitchell, P H, *The Measurement of Instrument Flying Proficiency of Air Force Pilots* Washington U S Government Printing Office, 1947

11 Jenkins, J G, Validity for What?" *Journal of Consulting Psychology*, 1946, Vol 10, 93–98

12 Johnson, H M, and Boots M L, *Analysis of Ratings in the Preliminary Phase of the CAA Training Program* Washington CAA Division of Research, Report No 21, 1943

13 Kelly E L, *The Development of "A Scale for Rating Pilot Competency"* Washington CAA Division of Research, Report No 18, 1943

14 Lepley, W M, *Psychological Research in the Theatres of War* Army Air Forces Aviation Psychology Program Research Report No 17 Washington U S Government Printing Office, 1947

15 McFarland R A, and Holway, A H, *The Measurement of Flight Performance in Relation to Piloting* Progress Report, National Research Council Committee on Aviation Psychology, 1942

16 McFarland, R A, and Holway, A H, *The Theory and Measurement of Flight Performance* Progress Report, National Research Council Committee on Aviation Psychology, 1941

17 McKay, W, *The Development of the C A A—N R C Flight Recorder* Washington CAA Airman Development Division, Report No 35, 1944

18 Miller, N E (ed), *Psychological Research on Pilot Training* Army Air Forces Aviation Psychology Research Report No 8 Washington U S Government Printing Office, 1947

19 National Research Council Committee on Selection and Training of Aircraft Pilots *History and Development of the Ohio State Flight Inventory Part II Recent Versions and Current Applications* Washington CAA Division of Research, Report No 51, 1945

20 *Pilot Assessment During Elementary Flying Training*, Empire Central Flying School, Royal Air Force, Report No 18, 1946

21 Preston, H O, *Analysis of C A A Records on Airline Transport Pilots* Washington CAA Division of Research Report No 72, 1947

22 *Transport Command Categorization of Flying Personnel* Royal Air Force, 1948

23 Viteles, M S, *The Aircraft Pilot 5 Years of Research A Summary of Outcomes* Washington CAA Division of Research, Report No 46, 1945

24 Viteles, M S, and Backstrom O, Jr, *An Analysis of Graphic Records of Pilot Performance Obtained by Means of the R S Ride Recorder Part I* Washington CAA Division of Research, Report No 23 1943

25 Viteles, M S, Franzen R, and Rogers, R C *The Association between Ratings on Specific Maneuvers and Success or Failure in Flight Training of RAF Cadets* Washington CAA Airman Development Division, Report No 37, 1944

26 Viteles, M S, and Thompson, A S, *An Analysis of Photographic Records of Aircraft Pilot Performance* Washington CAA Division of Research, Report No 31, 1944

27 Viteles M S, and Thompson, A S, *The Use of Standard Flights and Motion Photography in the Analysis of Aircraft Pilot Performance* Washington CAA Division of Research, Report No 15, 1943

28 Walker, R Y, Wapner, S, Bakan, D, and Ewart, E S, *The Agreement Between Inspectors' Observations as Recorded on the Ohio State Flight Inventory and Instrument Readings Obtained from Photographic Records* Washington CAA Division of Research, Report No 67, 1946

29 Wapner, S, Festinger, L, and Odbert, H S, *Comparison of Student Pilot Performance on Successive Check Flights as Measured by Photographic Records* Washington CAA Division of Research, Report No 59 1946

30 Wherry, R J , and Rogers, R C , *A Factor Analysis of the Purdue Scale for Rating Pilot Competency"* Washington CAA Division of Research, Report No 18, 1943

31 Wickert, F (ed), *Psychological Research on Problems of Redistribution* Army Air Forces Aviation Psychology Research Report No 14 Washington U S Government Printing Office, 1947

32 Williams, A C , Jr , MacMillan, J W , and Jenkins, J G , *Preliminary Experimental Investigations of Tension as a Determinant of Performance in Flight Training* Washington CAA Division of Research, Report No 64, 1946

Index